`QuickSelect` Process and `QuickVal` Residual Convergence

by

Jason Matterer

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

September, 2015

Abstract

We define a sequence of tree-indexed processes closely related to the operation of the `QuickSelect` search algorithm (also known as `Find`) for all the various values of n (the number of input keys) and m (the rank of the desired order statistic among the keys). As a "master theorem" we establish convergence of these processes in a certain Banach space, from which known distributional convergence results as $n \to \infty$ about

(1) the number of key comparisons required

are easily recovered

(a) when $m/n \to \alpha \in [0, 1]$, and

(b) in the worst case over the choice of m.

From the master theorem it is also easy, for distributional convergence of

(2) the number of symbol comparisons required,

both to recover the known result in the case (a) of fixed quantile α and to establish our main new result in the case (b) of worst-case `Find`.

ABSTRACT

Our techniques allow us to unify the treatment of cases (1) and (2) and indeed to consider many other cost functions as well. Further, all our results provide a stronger mode of convergence (namely, convergence in L^p or almost surely) than convergence in distribution. Extensions to `MultipleQuickSelect` are discussed briefly.

Under mild assumptions on a general class of cost functions, Fill and Nakama [12] proved that the scaled cost of `QuickVal` $S_n^{(V)}/n$ converges in L^p and almost surely to a limit random variable S. For a general cost function, we consider what we term the `QuickVal` residual:

$$R_n^{(\mathrm{V})} := \frac{S_n^{(\mathrm{V})}}{n} - S.$$

The residual is of natural interest (especially in light of the previous analogous work on `QuickSort` [30, 16, 18, 34]). In the case of `QuickMin` with key-comparisons cost, we are able to calculate—à la Bindjeme and Fill for `QuickSort` [5]—the exact (and asymptotic) L^2-norm of the residual. We take the result as motivation for the scaling factor $\sqrt{n}$ for the `QuickVal` residual for *general* population quantiles and for *general* cost. We then show *in general* that $\sqrt{n}\, R_n^{(\mathrm{V})}$ converges in law to a scale-mixture of centered Gaussians. In the motivating case of `QuickMin` with key-comparisons cost, we also prove convergence of moments. Following Fuchs for `QuickSort` [16], this was part of our original attempt at proving convergence in distribution for $\sqrt{n}\, R_n^{(\mathrm{V})}$: using the method of moments. However, this approach encountered two roadblocks: first, because the approach is inherently recursive, we don't know how to extend beyond the motivating case; second, in the case of `QuickMin` with key-comparisons cost, we can

ABSTRACT

show via Krein's condition [1] that the moments of the limiting distribution do *not* uniquely define a distribution . In the case of `QuickVal` with general cost functions, it is still an open question whether the limiting residual distribution is uniquely defined by its moments for any value of α.

Primary Reader: James A. Fill

Secondary Reader: John C. Wierman

Acknowledgments

Writing a dissertation is a long process full of ups and downs. What I have accomplished would not have been possible if it were not for the constant support of everyone around me. I could not have asked for a better advisor in Professor Jim Fill, whose insight and patience proved invaluable throughout the research process. If I were to try to list what I have learned from Professor Fill over the years from mathematical results down to typographical practices that would entail a dissertation of its own.

I owe a debt of gratitude to Emily Cahill, my girlfriend, for always being there to listen to my endless rambling about the work that makes up this dissertation. Without the nourishing support from my parents, I would never have developed my love of mathematics which makes all this possible.

Contents

Chapter 1

Introduction

In this dissertation, we analyze distributional asymptotics of the randomized search algorithm `QuickSelect`. In the first half of this dissertation, we define a sequence of tree-indexed processes closely related to the operation of the `QuickSelect` search algorithm (also known as `Find`) for all the various values of n (the number of input keys) and m (the rank of the desired order statistic among the keys). As a "master theorem" we establish convergence of these processes in a certain Banach space, from which known distributional convergence results as $n \to \infty$ about the number of key or symbol comparisons are recovered both (a) when $m/n \to \alpha \in [0, 1]$, and (b) in the worst case over the choice of m. In the second half, we consider an algorithm related to `QuickSelect` introduced in [35] called `QuickVal`. Motivated by previous limiting results from Fill and Nakama [12] and L^2 asymptotics [12] we prove in Chapter 5, we define what we call the `QuickVal` *residual* and prove convergence in distribution to a scale-mixture of centered Gaussians in Chapter 6. We also highlight an application of this result to `QuickMin` and `QuickMax`.

`QuickSelect` (also known as `Find`), introduced by Hoare [20], is a randomized algorithm for selecting a specified order statistic from an input sequence of objects, or rather their identifying labels usually known as *keys*. The keys can be numeric or symbol strings, or indeed any labels drawn from a given linearly ordered set. Suppose we are given keys $y_1, \ldots, y_n$ and we want to find the m^{th} smallest among them. The algorithm first selects a key (called the pivot) uniformly at random. It then compares every other key to the pivot, thereby determining the rank, call it r, of the pivot among the n keys. If $r = m$, then the algorithm terminates, returning the pivot key as output. If $r > m$, then the algorithm is applied recursively to the keys smaller than the pivot to find the mth smallest among those; while if $r < m$, then the algorithm is applied recursively to the keys larger than the pivot to find the $(m - r)$th smallest among those. More formal descriptions of `QuickSelect` can be found in [20] and [24], for example. When the desired rank equals 1, so that `QuickSelect` is searching for the minimum of the keys, we will use the name `QuickMin`.

When considering asymptotics of the cost of `QuickSelect` as the number of keys tends to infinity, it becomes necessary to let the order statistic m_n depend on the number of keys n. When $m_n/n \to \alpha$, we refer to `QuickSelect` finding the m_n^{th} order statistic in n keys as `QuickQuant`(n, α).

The `QuickVal` algorithm [35] is useful in the analysis of asymptotic costs of

`QuickQuant`. For n keys, `QuickVal`(n, α) searches for the α-quantile of the *population* from which the n keys are drawn. On the other hand, `QuickQuant`(n, α) finds the (approximate) α-quantile from the *sample* of n keys. A detailed description of `QuickVal` is postponed until we discuss the probabilistic sources of the keys (see Section 6.1).

Observe that, for fixed n and a given sequence $(y_1, \ldots, y_n)$ of keys, it is possible to build the randomness needed to run `QuickSelect` for *every* value of $m \in \{1, \ldots, n\}$ on a single probability space, as follows. Let π denote a uniformly random permutation of $\{1, \ldots, n\}$, and consider the sequence $(z_1, \ldots, z_n)$ with $z_i := y_{\pi_i}$ for $i = 1, \ldots, n$. Regardless of the value of m, choose z_1 as the initial pivot; and when the algorithm is applied recursively, apply it to the appropriate sequence of z_i-values listed in the *same* relative order as within $(z_1, \ldots, z_n)$.

The cost of running `QuickSelect` can be measured by assessing the cost of comparing keys. We assume that every comparison of two (distinct) keys costs some amount that is perhaps dependent on the values of the keys, and then the cost of the algorithm is the sum of the comparison costs.

Until recently, it has been customary to assign unit cost to each comparison of two keys, irrespective of their values. We denote the (random) key-comparisons-count cost for `QuickSelect` by $K_{n,m}$. As we have explained, for fixed n one can use a single uniformly random permutation of $\{1, \ldots, n\}$ to build a single probability space on which all of the random variables $K_{n,m}$ with $1 \leq m \leq n$ are defined. [Note also that the joint distribution of $K_{n,1}, \ldots, K_{n,n}$ does not depend on the initial sequence $(y_1, \ldots, y_n)$ of distinct keys.] Among other things, this opens up the possibility of studying the distribution of $\max_m K_{n,m}$, the cost of so-called "worst-case `Find`", in which an adversary is allowed to choose the rank of the key sought by the `QuickSelect` algorithm. Our motivation for the first half of this dissertation was to investigate the large-n behavior of worst-case `Find` for more general cost functions.

There have been many studies of the random variables $K_{n,m}$, including [8], [27], [19], [25], [17], [9], [22], [10], and [14]; and several corresponding studies, including [31], [26], and [28], of the number(s) of key comparisons for an extension of `QuickSelect` called `MultipleQuickSelect` that searches simultaneously for multiple order statistics. Grübel and Rösler [19] analyzed a modified version of `QuickSelect` that splits the collection of keys into two sets, those smaller than the pivot and those greater than or equal to the pivot, rather than into three sets (one of which has the pivot as its only element) as considered in this dissertation. They studied (see especially their Theorem 4) the limiting behavior of this modified `QuickSelect` through the convergence (in distribution, in the Skorohod topology on the space $D[0,1]$ of *càdlàg* functions on the unit interval $[0,1]$) of a sequence $X_1, X_2, \ldots$ of stochastic processes defined by $X_n(\alpha) := n^{-1} K_{n,\lfloor n\alpha \rfloor + 1}$ for $\alpha \in [0,1)$ and $X_n(1) := n^{-1} K_{n,n}$. Rüschendorf [25, Examples 4.1–4.2] utilized the contraction method to prove that the scale-normalized key-comparisons-count cost $n^{-1} \max_m K_{n,m}$ of worst-case `Find` (the version considered in this dissertation) converges in distribution. Devroye [9] presented an alternative proof of the latter result.

But unit cost is not always a reasonable model for comparing two keys. For example, if each key is a string of symbols, then a more realistic model for the cost of comparing two keys is the value of the first index at which the two symbol strings differ. To date, only a few papers ([35], [11], and [12]) have considered `QuickSelect` from this more realistic symbol-comparisons perspective. As in [12], in this dissertation we will treat a rather general class of cost functions that includes both key-comparisons cost and symbol-comparisons cost.

In our set-up (to be described in detail in Chapter 2) for this dissertation, we will consider a variety of probabilistic models (called *probabilistic sources*) for how a key is generated as an infinite-length string of symbols, but we will always assume that the keys form an infinite sequence of independent and identically distributed and almost surely distinct symbol strings. This gives us, on a single probability space, all the randomness needed to run `QuickSelect` for *every* value of n and *every* value of $m \in \{1, \ldots, n\}$ by always choosing the *first* key in the sequence as the pivot (and maintaining initial relative order of keys when the algorithm is applied recursively); this is what is meant by the *natural coupling* (cf. [13, Section 1]) of the runs of the algorithm for varying n and m. As explained in [13, Section 1], the coupling allows us to consider stronger forms of convergence than convergence in distribution, such as almost sure convergence and convergence in L^p.

Whatever cost function is used for comparisons of two keys, denote the corresponding total cost of `QuickSelect` (under the natural coupling) in selecting the mth order statistic from the first n keys by $\texttt{FIND}(n, m)$. Let $m_n \in \{1, \ldots, n\}$ for every n, and suppose that $m_n/n \to \alpha \in [0, 1]$. Fill and Nakama [12] prove, under certain "tameness" conditions (to be reviewed later) on the probabilistic source and the cost function, that $n^{-1}\texttt{FIND}(n, m_n)$ converges both in L^p and almost surely to a limiting random variable. We complement their result by proving analogous results for the cost of worst-case `Find`, namely, $\max_{1 \leq m \leq n} \texttt{FIND}(n, m)$. Our new results and (under somewhat stronger hypotheses than assumed in [12]) the results of Fill and Nakama [12] are both obtained rather effortlessly from a "master theorem", Theorem 3.1, which establishes convergence in a certain Banach space of a certain sequence of tree-indexed processes closely related to the operation of `QuickSelect` for all the various values of n and m.

The ubiquitous `QuickSort`, also introduced by Hoare [21], is a divide-and-conquer algorithm designed, in the same style as `QuickSelect`, to sort a collection of values efficiently. We will describe the algorithm as acting on an array of data; however note that the algorithm is data-representation agnostic. Given an array of n values (termed *keys*, as in `QuickSelect`), if $n = 0$ the algorithm returns the empty list (trivially sorted) and otherwise a pivot is chosen uniformly at random. The pivot is compared to the remaining $n - 1$ keys and the array is divided into the subarray of keys smaller than the pivot (say $X_<$) and subarray of keys larger than the pivot (say $X_>$). The keys are rearranged so that the entire subarray $X_<$ is to the left of the pivot and $X_>$ lies to the right. This function is then recursively called on both subarrays $X_<$ and $X_>$. A more complete description can be found in any algorithms textbook (e.g., [7]).

Régnier [32] proved almost sure convergence of the key-comparison cost of `QuickSort` to a limit random variable. Using the contraction method, Neininger [30] proved that the suitably normalized difference between the suitably normalized key-comparison cost of `QuickSort` and its almost sure limit converges in law to a standard normal distribution. Fuchs [16] provided an alternate proof of Neininger's result using the method of moments. Grübel and Kabluchko [18] extend Neininger and Fuch's result to what they call almost sure weak convergence. They prove a functional central limit theorem for the Biggins martingale [3] on the branching random walk and apply this result to `QuickSort` by appealing to the binary search tree characterization of `QuickSort` (we state a similar characterization for `QuickSelect` in Section 2.2).

We follow in the footsteps of Neininger and Fuchs and prove a similar limit law for `QuickVal`. Our technique applies the classical central limit theorem to an approximation to the cost of `QuickVal`, where we condition on a fixed number K of initial pivots. We then prove a series of reductions, using results like Slutsky's theorem (e.g., [2, Theorem A.14.9]), until we reach the desired convergence.

An outline for this dissertation is as follows. First, in Chapter 2 we carefully describe our set-up and, in some detail, discuss probabilistic sources, cost functions, and tameness; we also discuss the idea of *seeds*, which allow us a unified treatment of all sources. In Section 2.3 we state and prove a number of useful lemmas.

In Chapter 3 (specifically, our master Theorem 3.1) we prove that the "`QuickSelect` tree processes" to which we alluded earlier converge in a certain Banach space described in Definition 2.13 and Proposition 2.14. Some consequences of Theorem 3.1 are provided in Chapter 4; the highlight is Corollary 4.4, which gives sufficient conditions for L^p-convergence of the cost of worst-case `Find`. In Section 4.1, we use Theorem 3.1 to provide (under an additional restriction) very simple proofs of Theorems 3.1 and 4.1 in [12]; the latter concerns L^p-convergence of the cost of `QuickSelect` for fixed α. In Section 4.2 we complement the L^p-convergence result of Corollary 4.4 for the cost of worst-case `Find` by providing a tameness condition under which the scale-normalized cost of worst-case `Find` converges almost surely.

We motivate our scaling of `QuickVal` residual by $\sqrt{n}$ in Chapter 5 by first calculating the exact L^2 distance between the normalized key-comparison cost of `QuickMin` and its almost sure limit random variable. In Chapter 6, we consider the residual `QuickVal` $R_n^{(\mathrm{V})}$ random variable. Our work there culminates in Theorem 6.4 in Section 6.2 giving conditions for convergence in law of the `QuickVal` residual to a scale mixture of centered Gaussians. We approach the proof by proposing approximations to $S_n^{(\mathrm{V})}$ and S that consider only a fixed number of pivots (as opposed to letting the number of pivots considered tend to infinity as $n \to \infty$). We prove convergence of the `QuickVal` residual in Section 6.2 through an application of the classical central limit theorem and a series of reductions. Observing that the operation of `QuickMin` is equivalent to `QuickVal` when $\alpha = 0$, we immediately conclude Corollary 6.5. In Section 6.3, we prove using similar arguments used in the proof of Theorem 6.4 that the moments of $R_n^{(\mathrm{V})}$ converge to the moments of the limit scale mixture of centered Gaussians.

In Chapter 7 we consider the moments of the `QuickMin` residual and prove they converge to moments of a distribution defined by a distributional fixed point equation, which has as its unique solution a scale mixture of centered Gaussians. A recurrence for moments of the candidate limit distribution is derived in Section 7.3. In Section 7.2, we prove that a scale-mixture of centered Gaussians is the unique solution to the distributional fixed point equation. In Section 7.4, we derive a recurrence for the moments of the `QuickMin` residual under key-comparison cost, and prove that each recurrence converges (in a suitable sense) to the recurrence derived in Section 7.3.

Remark 1.1. As we recall from [12] at the end of our Section 2.1, many common sources, including memoryless and Markov sources, have the property that the source-specific cost function β corresponding to the symbol-comparisons cost for comparing keys is ϵ-tame for every $\epsilon > 0$. Thus, for such sources, the conclusions of two of our main results, Theorem 3.1 and Corollary 4.4, hold for every $p \in [2, \infty)$; and the almost-sure convergence theorem (Theorem 4.9) for worst-case `Find` and `QuickVal` residual convergence Theorem 6.4 also applies to all such sources.

Chapter 2

Set-up and Preliminaries

2.1 Probabilistic sources

Let us define the fundamental probabilistic structure underlying the analysis of `QuickSelect`. We assume that keys arrive independently and with the same distribution and that each key is composed of a sequence of symbols from some finite or countably infinite alphabet. Let Σ be this alphabet (which we assume is totally ordered by $\leq$). Then a key is an element of Σ^∞ [ordered by the lexicographic order, call it $\preceq$, corresponding to $(\Sigma, \leq)$] and a *probabilistic source* is a stochastic process $W = (W_1, W_2, W_3, \ldots)$ such that for each i the random variable W_i takes values in Σ. We will impose restrictions on the distribution of W that will have as a consequence that (with probability one) all keys are distinct.

We denote the cost (assumed to be nonnegative) of comparing two keys w, w'

by $\mathrm{cost}(w, w')$. As two examples, the choice $\mathrm{cost}(w, w') \equiv 1$ gives rise to a key-comparisons analysis, whereas if words are symbol strings then a symbol-comparisons analysis is obtained by letting $\mathrm{cost}(w, w')$ be the first index at which w and w' disagree.

Since Σ^∞ is totally ordered, a probabilistic source W is governed by a distribution function F defined for $w \in \Sigma^\infty$ by

$$F(w) := \mathbb{P}(W \preceq w).$$

Then the corresponding inverse probability transform M, defined by

$$M(u) := \inf \{w \in \Sigma^\infty : u \leq F(w)\},$$

has the property that if $U \sim \mathrm{uniform}(0, 1)$, then $M(U)$ has the same distribution as W. We refer to such uniform random variables U as *seeds*.

Using this technique we can define a *source-specific cost function*

$$\beta : (0, 1) \times (0, 1) \to [0, \infty)$$

by $\beta(u, v) := \mathrm{cost}(M(u), M(v))$.

Definition 2.1. Let $0 < c < \infty$ and $0 < \epsilon < \infty$. A source-specific cost function β is said to be (c, ϵ)-*tame* if for $0 < u < t < 1$ we have

$$\beta(u, t) \leq c\,(t - u)^{-\epsilon},$$

and is said to be ϵ-*tame* if it is (c, ϵ)-tame for some c.

For further important background on sources, cost functions, and tameness, we refer the reader to Section 2.1 (see especially Definitions 2.3–2.4 and Remark 2.5) in Fill and Nakama [12]. Note in particular that many common sources, including memoryless and Markov sources, have the property that the source-specific cost function β corresponding to symbol-comparisons cost for comparing keys is ϵ-tame for every $\epsilon > 0$.

2.2 Tree of seeds and the `QuickSelect` tree processes

Let $\mathcal{T}$ be the collection of (finite or infinite) rooted ordered binary trees (whenever we refer to a binary tree we will assume it is of this variety) and let $\overline{T} \in \mathcal{T}$ be the complete infinite binary tree. We will label each node θ in such a tree by a binary sequence representing the path from the root to θ, where 0 corresponds to taking the left child and 1 to taking the right. We consider the set of real-valued stochastic processes each with index set equal to some $T \in \mathcal{T}$. For such a process, we extend the index set to $\overline{T}$ by defining $X_\theta = 0$ for $\theta \in \overline{T} \setminus T$. This convention allows us to define addition of any two such processes componentwise, as well as scalar multiplication componentwise. In doing so, we obtain a vector space $\mathcal{B}$ of such processes. We will have need for the following definition of levels of a binary tree.

Definition 2.2. For $0 \le k < \infty$, we define the k^{th} *level* Λ_k of a binary tree as the collection of vertices that are at distance k from the root.

Let $\Theta = \bigcup_{0 \le k < \infty} \{0,1\}^k$ be the set of all finite-length binary strings, where $\{0,1\}^0 = \{\varepsilon\}$ with ε denoting the empty string. Set $L_\varepsilon := 0$, $R_\varepsilon := 1$, and $\tau_\varepsilon := 1$. Then, for $\theta \in \Theta$, we define $|\theta|$ to be the length of the string θ, and $\upsilon_\theta(n)$ to be the size (through the arrival of the n^{th} key) of the subtree rooted at node θ. Given a sequence of independent and identically distributed (iid) seeds $U_1, U_2, U_3, \ldots$, we recursively define

$$\begin{aligned} \tau_\theta &:= \inf\{i : L_\theta < U_i < R_\theta\}, \\ L_{\theta 0} &:= L_\theta, \ L_{\theta 1} := U_{\tau_\theta}, \\ R_{\theta 0} &:= U_{\tau_\theta}, \ R_{\theta 1} := R_\theta, \end{aligned}$$

where $\theta_1 \theta_2$ denotes the concatenation of $\theta_1, \theta_2 \in \Theta$. For a source-specific cost function β and $0 \le p < \infty$ we define

$$\begin{aligned} S_{\theta,n} &:= \sum_{\tau_\theta < i \le n} \mathbf{1}(L_\theta < U_i < R_\theta) \beta(U_i, U_{\tau_\theta}), \\ I_p(x,a,b) &:= \int_a^b \beta^p(u,x)\, du, \\ I_{p,\theta} &:= I_p(U_{\tau_\theta}, L_\theta, R_\theta), \\ I_\theta &:= I_{1,\theta}, \\ C_\theta &:= (\tau_\theta, U_{\tau_\theta}, L_\theta, R_\theta). \end{aligned}$$

In some later definitions we will make use of the positive part function defined as usual by $x^+ := x\mathbf{1}(x > 0)$. Given a source-specific cost function β and the seeds $U_1, U_2, U_3, \ldots$, we define the nth `QuickSelect` seed process as the n-nodes binary tree indexed stochastic process obtained by successive insertions of $U_1, \ldots, U_n$ into an initially empty binary search tree.

Before we use these random variables, we supply some understanding of them for the reader. The arrival time τ_θ is the index of the seed that is slotted into node θ in the construction of the `QuickSelect` seed process. Note that for each $\theta \in \Theta$ we have $\mathbb{P}(\tau_\theta < \infty) = 1$. The interval (L_θ, R_θ) provides sharp bounds for all seeds arriving after time τ_θ that interact with U_{τ_θ} in the sense of being placed in the subtree rooted at U_{τ_θ}. A crucial observation is that, conditioned on C_θ, the sequence of seeds $U_{\tau_\theta+1}, U_{\tau_\theta+2}, \ldots$ are iid uniform$(0,1)$; thus, again conditioned on C_θ, the sum $S_{\theta,n}$ is the sum of $(n - \tau_\theta)^+$ iid random variables. Note that when $n \leq \tau_\theta$ the sum defining $S_{\theta,n}$ is empty and so $S_{\theta,n} = 0$; in this case we shall conveniently interpret $S_{\theta,n}/(n - \tau_\theta)^+ = 0/0$ as 0. The random variable $S_{\theta,n}$ is the total cost of comparing the key with seed U_{τ_θ} with keys (among the first n to arrive) whose seeds fall in the interval (L_θ, R_θ), and $I_{p,\theta}$ is the conditional pth moment of one such comparison: If we let $U \sim$ uniform$(0,1)$ independent of C_θ, then

$$I_{p,\theta} = \mathbb{E}\left[\mathbf{1}(L_\theta < U < R_\theta)\beta^p(U, U_{\tau_\theta}) \middle| \, C_\theta\right].$$

Conditioned on C_θ, the expression $S_{\theta,n}$ is the sum of $(n - \tau_\theta)^+$ iid random variables with pth moment $I_{p,\theta}$.

When considering `QuickVal`, we will simplify the notation since we will only need to reference one path of nodes from the root to a leaf in the `QuickSelect` process tree. For this we define similar notation indexed by the pivot index (that is, by the level in the tree). Set $L_0 := 0$, $R_0 := 1$, and $\tau_0 := 1$. Then, for $k \geq 1$, we define

$$\begin{aligned}
\tau_k &:= \inf\{i : L_{k-1} < U_i < R_{k-1}\}, && (2.1)\\
L_k &:= \mathbf{1}(U_{\tau_k} < \alpha)U_{\tau_k} + \mathbf{1}(U_{\tau_k} > \alpha)L_{k-1}, && (2.2)\\
R_k &:= \mathbf{1}(U_{\tau_k} < \alpha)R_{k-1} + \mathbf{1}(U_{\tau_k} > \alpha)U_{\tau_k}, && (2.3)\\
C_k &:= (L_{k-1}, R_{k-1}, \tau_k, U_{\tau_k}) && (2.4)\\
X_{k,i} &:= \mathbf{1}(L_{k-1} < U_i < R_{k-1})\beta(U_i, U_{\tau_k}), && (2.5)\\
S_{k,n} &:= \sum_{i:\, \tau_k < i \leq n} X_{k,i}. && (2.6)
\end{aligned}$$

Remark 2.3. Note that [12] used the notation $S_{n,k}$ for what we have called $S_{k,n}$.

The random variable τ_k is the arrival time/index of the k^{th} pivot. The interval (L_k, R_k) gives the range of seeds to be compared to the k^{th} pivot in the operation of the `QuickVal` algorithm. The cost of comparing seed i to the k^{th} pivot is given by $X_{k,i}$. The total comparison costs attributed to the k^{th} pivot is $S_{k,n}$.

We define the n^{th} `QuickSelect` *tree process* as the binary-tree-indexed stochastic process $S_n = (S_{n,\theta})_{\theta \in \Theta}$ and the *limit* `QuickSelect` *tree process* (so called in light of Theorem 3.1) by $I = (I_\theta)_{\theta \in \Theta}$.

2.3 Preliminaries

We first prove some elementary lemmas that will be integral to the arguments used in the remainder of the paper. An important technique that will prove effective will be to bound moments of $I_{s,\theta}$ where $\theta \in \Lambda_k$ by an expression with geometric decrease in k. The following lemma provides such a bound in the case of an ϵ-tame source.

Lemma 2.4. *If β is (c, ϵ)-tame with $0 \le \epsilon < 1/s$, then for each fixed node $\theta \in \Lambda_k$ and $0 \le r < \infty$ we have*

$$\mathbb{E} I^r_{s,\theta} \le \left(\frac{2^{s\epsilon} c^s}{1 - s\epsilon} \right)^r \left(\frac{1}{r + 1 - rs\epsilon} \right)^k.$$

Proof. By ϵ-tameness and concavity of the $(1 - s\epsilon)$-power function,

$$\begin{aligned} I_{s,\theta} &\le c^s \int_{L_\theta}^{R_\theta} |u - U_{\tau_\theta}|^{-s\epsilon} \, du \\ &= \frac{c^s}{1 - s\epsilon} \left[(R_\theta - U_{\tau_\theta})^{1-s\epsilon} + (U_{\tau_\theta} - L_\theta)^{1-s\epsilon} \right] \\ &\le \frac{2^{s\epsilon} c^s}{1 - s\epsilon} (R_\theta - L_\theta)^{1-s\epsilon}. \end{aligned}$$

Since $R_\theta - L_\theta$ is distributed as the product of k independent uniform$(0, 1)$ random variables, taking rth moments gives the desired bound. □

As a consequence of Lemma 2.4, we have

Lemma 2.5. *Let $1 \le p < \infty$ and consider a fixed node θ. If the source-specific cost function β is ϵ-tame for some $0 \le \epsilon < 1/p$, then as $n \to \infty$ we have*

$$\frac{S_{\theta,n}}{n} \xrightarrow{L^p} I_\theta.$$

Proof. The proof essentially repeats an argument within the proof of Theorem 3.1 in Fill and Nakama [12]. Conditioned on C_θ, the random variable $S_{\theta,n}$ is the sum of $(n-\tau_\theta)^+$ iid nonnegative random variables with expectation I_θ and pth moment $I_{p,\theta}$. The L^p law of large numbers (L^pLLN) applies almost surely because $\mathbb{E}I_{p,\theta} < \infty$ by Lemma 2.4 and hence $I_{p,\theta} < \infty$ almost surely. The L^pLLN gives

$$\mathbb{E}\left[\left|\frac{S_{\theta,n}}{(n-\tau_\theta)^+} - I_\theta\right|^p \middle| C_\theta\right] \to 0 \quad \text{a.s.} \tag{2.7}$$

By convexity of the pth-power function,

$$\mathbb{E}\left[\left|\frac{S_{\theta,n}}{(n-\tau_\theta)^+} - I_\theta\right|^p \middle| C_\theta\right] \leq 2^{p-1}\left\{\mathbb{E}\left[\left(\frac{S_{\theta,n}}{(n-\tau_\theta)^+}\right)^p \middle| C_\theta\right] + I_\theta^p\right\}$$

and also

$$\left(\frac{S_{n,\theta}}{(n-\tau_\theta)^+}\right)^p \leq \frac{1}{(n-\tau_\theta)^+}\sum_{i:\tau_\theta<i\leq n}\mathbf{1}(L_\theta < U_i < R_\theta)\beta^p(U_i, U_{\tau_\theta}),$$

which implies

$$\mathbb{E}\left[\left(\frac{S_{\theta,n}}{(n-\tau_\theta)^+}\right)^p \middle| C_\theta\right] \leq I_{p,\theta}.$$

Therefore, we have the following bound:

$$\mathbb{E}\left[\left|\frac{S_{\theta,n}}{(n-\tau_\theta)^+} - I_\theta\right|^p \middle| C_\theta\right] \leq 2^{p-1}\left(I_{p,\theta} + I_\theta^p\right) \leq 2^p I_{p,\theta}. \tag{2.8}$$

Recall that Lemma 2.4 implies that $\mathbb{E}I_{p,\theta} < \infty$. Therefore, by (2.7)–(2.8) and the dominated convergence theorem,

$$\mathbb{E}\left|\frac{S_{\theta,n}}{(n-\tau_\theta)^+} - I_\theta\right|^p \to 0.$$

Now to complete the proof of the lemma we show that

$$\mathbb{E}\left|\frac{S_{\theta,n}}{n}-\frac{S_{\theta,n}}{(n-\tau_\theta)^+}\right|^p \to 0.$$

By our convention for $S_{n,\theta}/(n-\tau_\theta)^+$ when $n \leq \tau_\theta$, we have

$$\left|\frac{S_{\theta,n}}{n}-\frac{S_{\theta,n}}{(n-\tau_\theta)^+}\right|^p = \left|\frac{S_{\theta,n}}{n}-\frac{S_{\theta,n}}{(n-\tau_\theta)^+}\right|^p \mathbf{1}(\tau_\theta < n).$$

By a simple calculation,

$$\left|\frac{S_{\theta,n}}{n}-\frac{S_{\theta,n}}{(n-\tau_\theta)^+}\right|^p = \left|\frac{\tau_\theta}{n}\right|^p \left|\frac{S_{\theta,n}}{(n-\tau_\theta)^+}\right|^p.$$

Taking expectations conditioned on C_θ gives

$$\begin{aligned}\mathbb{E}\left[\left|\frac{S_{\theta,n}}{n}-\frac{S_{\theta,n}}{(n-\tau_\theta)^+}\right|^p \mathbf{1}(\tau_\theta < n)\right| C_\theta\right] &= \mathbf{1}(\tau_\theta < n)\left(\frac{\tau_\theta}{n}\right)^p \mathbb{E}\left[\left|\frac{S_{\theta,n}}{(n-\tau_\theta)^+}\right|^p \middle| C_\theta\right] \\ &\leq \mathbf{1}(\tau_\theta < n)\left(\frac{\tau_\theta}{n}\right)^p I_{p,\theta} \leq I_{p,\theta},\end{aligned}$$

and, in particular,

$$\mathbb{E}\left[\left|\frac{S_{\theta,n}}{n}-\frac{S_{\theta,n}}{(n-\tau_\theta)^+}\right|^p \middle| C_\theta\right] \to 0 \quad \text{a.s.}$$

Thus, again by the dominated convergence theorem,

$$\mathbb{E}\left|\frac{S_{\theta,n}}{n}-\frac{S_{\theta,n}}{(n-\tau_\theta)^+}\right|^p \to 0.$$

□

A Poisson binomial sum is a generalization of a binomial distributed random variable. Let $X_i \sim \mathrm{Bern}(p_i)$, $i = 1, 2, \ldots, n$, be independent, where $\mathrm{Bern}(p)$ denotes the Bernoulli distribution with success probability p. Then we say that $X := \sum_i X_i$

is a *Poisson binomial sum.* The following lemma is a restatement of Theorem 4.4 (part 1) in [29]; we include a proof for completeness.

Lemma 2.6. *Let* $X = \sum_i X_i$ *be a Poisson binomial sum with* $X_i \sim \mathrm{Bern}(p_i)$ *and* $\mathbb{E}X = \sum_i p_i =: \mu$. *Then for any* $\beta \geq 0$ *we have*

$$\mathbb{P}(X \geq (1+\beta)\mu) \leq \left[\frac{e^{\beta}}{(1+\beta)^{1+\beta}}\right]^{\mu}.$$

Proof. The result is trivial for $\beta = 0$, so suppose $\beta > 0$. For any $t > 0$, by Markov's inequality

$$\mathbb{P}(X \geq (1+\beta)\mu) = \mathbb{P}(e^{tX} \geq e^{t(1+\beta)\mu}) \leq e^{-t(1+\beta)\mu}\mathbb{E}e^{tX}.$$

Since $X = \sum_i X_i$ and the X_i's are independent,

$$\mathbb{E}e^{tX} = \prod_{i=1}^{n} \mathbb{E}e^{tX_i} = \prod_{i=1}^{n} \left[1 + p_i\left(e^t - 1\right)\right] \leq \prod_{i=1}^{n} \exp\left[p_i\left(e^t - 1\right)\right] = \exp\left[\mu\left(e^t - 1\right)\right].$$

Combining these two inequalities and choosing $t = \ln(1+\beta) > 0$ produces the desired bound. □

Lemma 2.7. *Let* $H_m := \sum_{i=1}^{m} i^{-1}$, *the* mth *harmonic number. If* $\theta \in \Lambda_k$ *with* $k > H_m$, *then*

$$\mathbb{P}(\tau_\theta \leq m) \leq \left(\frac{eH_m}{k}\right)^k e^{-H_m}.$$

Proof. By symmetry, it suffices to consider $\theta = 0^k$ (the leftmost node of the binary tree at level k). Then τ_θ is the arrival time of the k^{th} record-smallest seed. If R_m is the number of seeds among the first m to be record-smallest upon arrival, then we have $\mathbb{P}(\tau_\theta \le m) = \mathbb{P}(R_m \ge k)$. It is well known that R_m has the distribution of a Poisson binomial sum:

$$R_m \stackrel{\mathcal{L}}{=} \sum_{i=1}^{m} X_i,$$

where $X_i \sim \text{Bern}(1/i)$, $i = 1, \ldots, m$, are independent and $\stackrel{\mathcal{L}}{=}$ denotes equality in law. (Consult, for example, [4, Problem 20.9].) This implies that $\mu := \mathbb{E}R_m = \sum_{i=1}^{m} i^{-1} = H_m$.

Thus for $\beta = (k/H_m) - 1 > 0$, applying Lemma 2.6 gives

$$\begin{aligned} \mathbb{P}(\tau_\theta \le m) = \mathbb{P}(R_m \ge k) &= \mathbb{P}(R_m \ge (1+\beta)\mu) \\ &\le \left[\frac{e^\beta}{(1+\beta)^{1+\beta}}\right]^\mu = \left(\frac{eH_m}{k}\right)^k e^{-H_m}. \end{aligned}$$

□

In the context of Lemma 2.7, we will use the following standard bound on harmonic numbers to approximate H_n by $\ln n$ up to a constant term.

Lemma 2.8. *Let γ be Euler's constant. Then for $n = 1, 2, \ldots$ we have*

$$\gamma \le H_n - \ln n \le 1.$$

The next lemma uses Lemma 2.4 in its proof and can be seen to generalize Lemma 2.4 (by letting $a \to 0$ and then $w \to 1$).

Lemma 2.9. *Consider a fixed node $\theta \in \Lambda_k$, and let $0 \le r < \infty$ and $0 < a < \infty$ be constants. Let the source-specific cost function β be (c, ϵ)-tame with $0 \le \epsilon < 1/s$.*

Then for any $v, w \in (1, \infty)$ such that $v^{-1} + w^{-1} = 1$, we have

$$\mathbb{E}\left(\tau_\theta{}^{-a} I^r_{s,\epsilon}\right) \leq \left(\frac{2^{s\epsilon} c^s}{1 - s\epsilon}\right)^r \left(\frac{1}{wr + 1 - wrs\epsilon}\right)^{k/w} \times \left\{ \left(2vak \int_1^{e-\delta_k} \left[\frac{e \ln \alpha + \delta_k}{\alpha^{1+va}}\right]^k d\alpha\right)^{1/v} + (e - \delta_k)^{-ak} \right\}$$

where $\delta_k = e/k$.

Proof. By Hölder's inequality, $\mathbb{E}\left(\tau_\theta^{-a} I^r_{s,\theta}\right) \leq \left\|\tau_\theta^{-a}\right\|_v \left\|I^r_{s,\theta}\right\|_w$. We can use Lemma 2.4 to bound the second factor. To treat the first factor, we express the expectation as an integral of tail probabilities, which we bound using Chernoff's inequality. Write

$$\begin{aligned} \mathbb{E}\tau_\theta^{-va} &= \int_0^\infty \mathbb{P}(\tau_\theta^{-va} \geq t)\, dt \\ &= \int_0^\infty \mathbb{P}(\tau_\theta \leq \lfloor t^{-1/(va)} \rfloor)\, dt := J. \end{aligned}$$

Using the change-of-variables $\alpha^k = t^{-1/(va)}$, we get

$$J = \int_0^\infty \mathbb{P}(\tau_\theta \leq \lfloor \alpha^k \rfloor)(vak)\alpha^{-vak-1}\, d\alpha. \tag{2.9}$$

The next step is to use Lemma 2.7 to bound $\mathbb{P}(\tau_\theta \leq \lfloor \alpha^k \rfloor)$. However, we need to check that the hypothesis of Lemma 2.7 that

$$k \geq H_{\lfloor \alpha^k \rfloor} \tag{2.10}$$

is satisfied. By Lemma 2.8 we have the upper bound

$$H_{\lfloor \alpha^k \rfloor} \leq \ln \lfloor \alpha^k \rfloor + 1 \leq k \ln \alpha + 1,$$

so $\alpha \leq e^{1-(1/k)}$ is sufficient for (2.10), as therefore is $\alpha \leq e[1 - (1/k)]$. Writing $\delta_k = e/k$, we decompose the integral $\int_0^\infty$ in (2.9) for J into $\int_0^1 + \int_1^{e-\delta_k} + \int_{e-\delta_k}^\infty$. When

$0 \le \alpha < 1$, we have $\mathbb{P}(\tau_\theta \le \lfloor \alpha^k \rfloor) = 0$ since $\tau_\theta \ge 1$. For the second integral we bound this probability using Lemma 2.7, and for the third integral we use the bound 1:

$$\begin{aligned} J &\le \int_1^{e-\delta_k} \left(\frac{eH_{\lfloor \alpha^k \rfloor}}{k} \right)^k e^{-H_{\lfloor \alpha^k \rfloor}} (vak)\alpha^{-vak-1}\, d\alpha + \int_{e-\delta_k}^{\infty} (vak)\alpha^{-vak-1}\, d\alpha \\ &= \int_1^{e-\delta_k} \left(\frac{eH_{\lfloor \alpha^k \rfloor}}{k} \right)^k e^{-H_{\lfloor \alpha^k \rfloor}} (vak)\alpha^{-vak-1}\, d\alpha + (e-\delta_k)^{-vak}. \end{aligned}$$

By Lemma 2.8,

$$\begin{aligned} \left(\frac{eH_{\lfloor \alpha^k \rfloor}}{k} \right)^k e^{-H_{\lfloor \alpha^k \rfloor}} &\le \left(\frac{e\left(\ln \lfloor \alpha^k \rfloor + 1\right)}{k} \right)^k e^{-\ln \lfloor \alpha^k \rfloor} \\ &\le \left[e \left(\ln \alpha + \frac{1}{k} \right) \right]^k \left(\frac{1}{\lfloor \alpha^k \rfloor} \right) \\ &\le 2 \left[\frac{e \ln \alpha + \delta_k}{\alpha} \right]^k, \end{aligned}$$

where in the last inequality we have used the fact that $2\lfloor x \rfloor \ge x$ for $x \ge 1$. Thus

$$\int_1^{e-\delta_k} \left(\frac{eH_{\lfloor \alpha^k \rfloor}}{k} \right)^k e^{-H_{\lfloor \alpha^k \rfloor}} (vak)\alpha^{-vak-1}\, d\alpha \le 2vak \int_1^{e-\delta_k} \left[\frac{e\ln\alpha + \delta_k}{\alpha^{1+va}} \right]^k d\alpha. \quad (2.11)$$

Bounding $\left\| I_{s,\theta}^r \right\|_w$ using Lemma 2.4 gives

$$\left\| I_{s,\theta}^r \right\|_w \le \left(\frac{2^{s\epsilon} c^s}{1 - s\epsilon} \right)^r \left(\frac{1}{wr + 1 - wrs\epsilon} \right)^{k/w}. \quad (2.12)$$

Combining (2.11) and (2.12) and using $(x+y)^{1/v} \le x^{1/v} + y^{1/v}$ for $x, y \ge 0$ proves the lemma. □

The following elementary calculus lemma will prove useful in the proof of Theorem 3.1.

Lemma 2.10. *The function $f(x) := ex^{-1}\ln x$ has a unique maximum for $x \in (0,\infty)$, at $\hat{x} = e$ with value $f(\hat{x}) = 1$.*

The following simple consequence of the triangle inequality will be used in the proof of Corollary 4.3.

Lemma 2.11. *Let $(x_i)_{i\in I}$ and $(y_i)_{i\in I}$ be two collections of nonnegative real numbers indexed by a common set I. Then*

$$\left|\sup_{i\in I} x_i - \sup_{i\in I} y_i\right| \leq \sup_{i\in I} |x_i - y_i|$$

provided at least one of the suprema on the left is finite (so that the difference of them is well defined).

Our next lemma is a simple bound, used in the proofs of Theorem 4.7 and Lemma 4.12, on the L^p-norm of a maximum of a finite collection of nonnegative random variables.

Lemma 2.12 (Max Lemma)**.** *Consider random variables $X_m \geq 0$, $m = 1, \ldots, M$, and let $1 \leq p < \infty$. Then*

$$\left\|\max_m X_m\right\|_p \leq M^{1/p} \max_m \|X_m\|_p .$$

Proof. The key to the proof is to bound the maximum of the M random variables X_m^p by their sum and then the sum of their pth moments by M times the maximum of the pth moments:

$$\begin{aligned}\left\|\max_m X_m\right\|_p^p &= \mathbb{E}\left(\max_m X_m\right)^p = \mathbb{E}\max_m X_m^p \\ &\leq \sum_m \mathbb{E}X_m^p \leq M \max_m \mathbb{E}X_m^p \\ &= M \max_m \|X_m\|_p^p .\end{aligned}$$

□

Recall from the first paragraph of Section 2.2 the definition of the vector space $\mathcal{B}$. Given any sequence of positive real numbers (a_k) and any $1 \le p \le \infty$, we can define a very useful functional on $\mathcal{B}$ as follows.

Definition 2.13. We define a real-valued function $|||\cdot|||_p$ on the vector space $\mathcal{B}$ of binary tree indexed stochastic processes (X_θ) by setting

$$|||X|||_p := \sum_{k=0}^{\infty} a_k \max_{\theta \in \Lambda_k} \|X_\theta\|_p .$$

Let $\mathcal{B}^{(p)} := \{X \in \mathcal{B} : |||X|||_p < \infty\}$. In typical fashion, we identify two processes X and Y in $\mathcal{B}^{(p)}$ if $|||X - Y|||_p = 0$.

Proposition 2.14. *$\mathcal{B}^{(p)}$ is a Banach space with the norm $|||\cdot|||_p$.*

Proof. It is easily verified that $\mathcal{B}^{(p)}$ is a vector subspace of $\mathcal{B}$ and that $|||\cdot|||_p$ is a norm on $\mathcal{B}^{(p)}$. What remains is to establish that the vector space $\mathcal{B}^{(p)}$ is complete with respect to the metric induced by $|||\cdot|||_p$. Our proof of this is adapted from Section 6.10 in [36]. Suppose $X^{(n)} \in \mathcal{B}^{(p)}$ form a Cauchy sequence with respect to the norm $|||\cdot|||_p$. Choose a sequence $\ell_n \to \infty$ as $n \to \infty$ such that for $m_1, m_2 \ge \ell_n$ we have

$$\left|\left|\left| X^{(m_1)} - X^{(m_2)} \right|\right|\right|_p \le 2^{-n}.$$

This implies that for each $\theta \in \Theta$, the random variables $X_\theta^{(n)}$ form a Cauchy sequence in L^p and for $m \ge \ell_n$ and $t \ge n$ we have

$$\left|\left|\left| X^{(m)} - X^{(\ell_t)} \right|\right|\right|_p = \sum_{k=0}^{\infty} a_k \max_{\theta \in \Lambda_k} \left\| X_\theta^{(m)} - X_\theta^{(\ell_t)} \right\|_p \le 2^{-n}.$$

For each θ, let X_θ be the L^p limit of $X_\theta^{(n)}$. For m fixed and $k = 1, 2, \ldots$, define

$$f_t(k) := \max_{\theta \in \Lambda_k} \left\| X_\theta^{(m)} - X_\theta^{(\ell_t)} \right\|_p,$$

and

$$f(k) := \max_{\theta \in \Lambda_k} \left\| X_\theta^{(m)} - X_\theta \right\|_p.$$

Since for every θ we have $X_\theta^{(n)} \xrightarrow{L^p} X_\theta$, for each k we have $f_t(k) \to f(k)$ as $t \to \infty$. Let μ be the counting measure induced by (a_k) on $\mathbb{Z}_+$, i.e., $\mu(\{k\}) = a_k$ for every $k \in \mathbb{Z}_+$. Then by Fatou's lemma,

$$\int \liminf_{t \to \infty} f_t \, d\mu \leq \liminf_{t \to \infty} \int f_t \, d\mu,$$

which, in the case of $m \geq \ell_n$, simplifies to

$$\left|\left|\left| X^{(m)} - X \right|\right|\right|_p \leq \liminf_{t \to \infty} \left|\left|\left| X^{(m)} - X^{(k_t)} \right|\right|\right|_p \leq 2^{-n},$$

where $X \in \mathcal{B}^{(p)}$ is the process of the X_θ's. Since this inequality holds for all $m \geq \ell_n$, we get $\left|\left|\left| X^{(m)} - X \right|\right|\right|_p \to 0$. □

We will need the solution to the following recurrence in Chapter 5 and Chapter 7 (specifically Theorem 5.1 and Proposition 7.9).

Lemma 2.15. *Let $\{(A_n)\}_{n \geq 0}$ and $\{(B_n)\}_{n \geq 1}$ be sequences of real numbers that satisfy*

$$A_n = \frac{1}{n} \sum_{k=0}^{n-1} A_k + B_n,$$

for $n \geq 1$. Then for $n \geq 0$ we have

$$A_n = A_0 + B_n + \sum_{k=1}^{n-1} \frac{1}{k+1} B_k, \tag{2.13}$$

with $B_0 := 0$.

Proof. The proof is by induction. When $n = 0$, we have $A_0 = A_0$ vacuously. Let $n \geq 1$ and assume (2.13) holds for $n-1$ and consider the recurrence for A_n multiplied by n. Then

$$\begin{aligned} nA_n &= nB_n + \sum_{k=0}^{n-1} A_k, \\ nA_n - (n-1)A_{n-1} &= nB_n - (n-1)B_{n-1} + A_{n-1}, \\ A_n &= B_n - \frac{n-1}{n} B_{n-1} + A_{n-1}. \end{aligned}$$

Applying the induction hypothesis gives

$$\begin{aligned} A_n &= A_0 + B_n - \frac{n-1}{n} B_{n-1} + B_{n-1} + \sum_{k=0}^{n-2} \frac{1}{k+1} B_k \\ &= A_0 + B_n + \frac{1}{n} B_{n-1} + \sum_{k=0}^{n-2} \frac{1}{k+1} B_k \\ &= A_0 + B_n + \sum_{k=0}^{n-1} \frac{1}{k+1} B_k. \end{aligned}$$

□

Chapter 3

Process convergence

Our main result is to show that the normalized `QuickSelect` tree process $S_n/n = (S_{n,\theta}/n)$ converges to a limit, namely, $I = (I_\theta)$, in the Banach space $(\mathcal{B}^{(p)}, |||\cdot|||_p)$ described in Proposition 2.14.

Theorem 3.1 (Master Theorem)**.** *For $2 \leq p < \infty$ and ϵ-tame source-specific cost function β with $0 \leq \epsilon < 1/p$, as $n \to \infty$ we have*

$$\left|\left|\left|\frac{S_n}{n} - I\right|\right|\right|_p \to 0,$$

with $a_k \equiv a_{k,p} \equiv 2^{k/p}$ in the definition of $|||\cdot|||_p$.

Proof. By Lemma 2.5, for each θ in the infinite binary tree, $\|(S_{\theta,n}/n) - I_\theta\|_p \to 0$, so that by the dominated convergence theorem it suffices to find a sequence (b_k) such that

(i) for $\theta \in \Lambda_k$ we have $\|(S_{\theta,n}/n) - I_\theta\|_p \leq b_k$, and

(ii) $\sum_{k=0}^{\infty} 2^{k/p} b_k < \infty$.

Decomposing I_θ into $I_\theta \mathbf{1}(\tau_\theta < n) + I_\theta \mathbf{1}(\tau_\theta \geq n)$ gives

$$\left| \frac{S_{\theta,n}}{n} - I_\theta \right| = \left| \left(\frac{S_{\theta,n}}{n} - \frac{n - \tau_\theta}{n} I_\theta \right) \mathbf{1}(\tau_\theta < n) - \frac{\tau_\theta}{n} I_\theta \mathbf{1}(\tau_\theta < n) - I_\theta \mathbf{1}(\tau_\theta \geq n) \right| \quad (3.1)$$

$$\leq \frac{1}{n} \left| S_{\theta,n} - (n - \tau_\theta) I_\theta \right| \mathbf{1}(\tau_\theta < n) + I_\theta, \quad (3.2)$$

so that

$$\left\| \frac{S_{\theta,n}}{n} - I_\theta \right\|_p \leq \left\| \frac{1}{n} \left| S_{\theta,n} - (n - \tau_\theta) I_\theta \right| \mathbf{1}(\tau_\theta < n) \right\|_p + \left\| I_\theta \right\|_p. \quad (3.3)$$

Recall that, conditionally given $C_\theta = (\tau_\theta, L_\theta, R_\theta, U_{\tau_\theta})$, the random variable $S_{\theta,n}$ is the iid sum of $(n - \tau_\theta)^+$ random variables with mean I_θ. If we let $U \sim$ uniform$(0,1)$ independent of C_θ and define $X_\theta := \mathbf{1}(L_\theta \leq U \leq R_\theta)\beta(U, U_{\tau_\theta})$, then by Rosenthal's inequality [33] there exists a constant c_p depending only on p such that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n^p} \left| S_{\theta,n} - (n - \tau_\theta) I_\theta \right|^p \mathbf{1}(\tau_\theta < n) \,\middle|\, C_\theta \right] \\
&\leq c_p \frac{1}{n^p} \mathbf{1}(\tau_\theta < n) \left\{ (n - \tau_\theta) \mathbb{E}\left[|X_\theta - I_\theta|^p \mid C_\theta \right] + (n - \tau_\theta)^{p/2} \left(\operatorname{Var}\left[X_\theta \,|C_\theta \right] \right)^{p/2} \right\} \\
&\leq c_p \mathbf{1}(\tau_\theta < n) \left\{ \frac{1}{n^{p-1}} \mathbb{E}\left[|X_\theta - I_\theta|^p \mid C_\theta \right] + \frac{1}{n^{p/2}} \left(\operatorname{Var}\left[X_\theta \,|C_\theta \right] \right)^{p/2} \right\}.
\end{aligned}$$

However, by convexity of the p^{th}-power function,

$$\mathbb{E}\left[|X_\theta - I_\theta|^p \,|C_\theta \right] \leq 2^{p-1} \left(\mathbb{E}\left[X_\theta^p \mid C_\theta \right] + I_\theta^p \right) = 2^{p-1} \left(I_{p,\theta} + I_\theta^p \right) \leq 2^p I_{p,\theta},$$

and $\operatorname{Var}[X_\theta | C_\theta] \leq I_{2,\theta}$. Thus, utilizing the factor $\mathbf{1}(\tau_\theta < n)$ gives

$$\mathbb{E}\left[\frac{1}{n^p} \left| S_{\theta,n} - (n - \tau_\theta) I_\theta \right|^p \mathbf{1}(\tau_\theta < n) \,\middle|\, C_\theta \right] \leq c_p 2^p \left(\tau_\theta^{-(p-1)} I_{p,\theta} + \tau_\theta^{-p/2} I_{2,\theta}^{p/2} \right).$$

Taking expectations yields

$$\left\| \frac{1}{n} \left| S_{\theta,n} - (n - \tau_\theta) I_\theta \right| \mathbf{1}(\tau_\theta < n) \right\|_p^p \\ \leq c_p 2^p \left(\mathbb{E}\left[\tau_\theta^{-(p-1)} I_{p,\theta} \right] + \mathbb{E}\left[\tau_\theta^{-p/2} I_{2,\theta}^{p/2} \right] \right). \tag{3.4}$$

From (3.3), (3.4), and Lemma 2.4 (with $s = 1$ and $r = p$), we get

$$\begin{aligned} &\left\| \frac{S_{\theta,n}}{n} - I_\theta \right\|_p^p \\ &\quad \leq \left[\left\| \frac{1}{n} \left| S_{\theta,n} - (n - \tau_\theta) I_\theta \right| \mathbf{1}(\tau_\theta < n) \right\|_p + \| I_\theta \|_p \right]^p \\ &\quad \leq 2^{p-1} \left[\left\| \frac{1}{n} \left| S_{\theta,n} - (n - \tau_\theta) I_\theta \right| \mathbf{1}(\tau_\theta < n) \right\|_p^p + \| I_\theta \|_p^p \right] \\ &\quad \leq 2^{p-1} \left[c_p 2^p \left(\mathbb{E}\left[\tau_\theta^{-(p-1)} I_{p,\theta} \right] + \mathbb{E}\left[\tau_\theta^{-p/2} I_{2,\theta}^{p/2} \right] \right) \right. \\ &\qquad\qquad \left. + \left(\frac{2^\epsilon c}{1 - \epsilon} \right)^p (p + 1 - p\epsilon)^{-k} \right]. \end{aligned} \tag{3.5}$$

Because $p + 1 - p\epsilon > p \geq 2$, it suffices to prove geometric decay at a rate faster than 2^{-k} for each of the two expectations in (3.5). (Note that when $p = 2$, the two expectations are equal, so we can and do restrict to $p > 2$ in considering the first expectation but allow $p = 2$ in considering the second.)

To establish this geometric decay, consider $\mathbb{E}\left[\tau_\theta^{-(p-1)} I_{p,\theta} \right]$ first. For any $v, w \in (1, \infty)$ satisfying $v^{-1} + w^{-1} = 1$, applying Lemma 2.9 gives

$$\begin{aligned} \mathbb{E}\left[\tau_\theta^{-(p-1)} I_{p,\theta} \right] &\leq \left(\frac{2^{p\epsilon} c^p}{1 - p\epsilon} \right) \left(\frac{1}{w + 1 - wp\epsilon} \right)^{k/w} \\ &\quad \times \left\{ \left(2v(p-1)k \int_1^{e-\delta_k} \left[\frac{e \ln \alpha + \delta_k}{\alpha^{1+v(p-1)}} \right]^k d\alpha \right)^{1/v} + (e - \delta_k)^{-(p-1)k} \right\}. \end{aligned}$$

The factor $(w+1-wp\epsilon)^{k/w}$ here is bounded by unity, and the term involving $(e-\delta_k)^{-(p-1)k}$ in the bound on $\mathbb{E}\left[\tau_\theta^{-(p-1)} I_{p,\theta}\right]$ decays at a geometric rate faster than 2^{-k}. So it suffices to show that there exist $v>1$ and $\eta>0$ [each of which may depend on p, but not on k or $\alpha \in (1, e-\delta_k)$] such that

$$2^v \left[\frac{e\ln\alpha + \delta_k}{\alpha^{1+v(p-1)}}\right] \le 1-\eta \tag{3.6}$$

for all large k. Since $\delta_k/\alpha^{1+v(p-1)} \le \delta_k \to 0$ as $k\to\infty$, it suffices to establish (3.6) with δ_k replaced by 0. By Lemma 2.10, we have

$$2^v\left(\frac{e\ln\alpha}{\alpha^{1+v(p-1)}}\right) \le \frac{2^v}{1+v(p-1)},$$

and the bound here is strictly less than 1 precisely when

$$p > 1 + \frac{2^v-1}{v}. \tag{3.7}$$

Since $1+v^{-1}(2^v-1)\to 2$ as $v\to 1$, for any $p>2$ a value of $v>1$ can be found satisfying (3.7).

Now consider $\mathbb{E}\left[\tau_\theta{}^{-p/2} I_{2,\theta}^{p/2}\right]$. Applying Lemma 2.9 gives

$$\mathbb{E}\tau_\theta{}^{-p/2} I_{2,\theta}^{p/2} \le \left(\frac{2^{2\epsilon}c^2}{1-2\epsilon}\right)^{p/2}\left(\frac{1}{(wp/2)+1-wp\epsilon}\right)^{k/w} \times\left\{\left(vpk\int_1^{e-\delta_k}\left[\frac{e\ln\alpha+\delta_k}{\alpha^{1+(vp/2)}}\right]^k d\alpha\right)^{1/v} + (e-\delta_k)^{-pk/2}\right\}.$$

As before, the term involving $(e-\delta_k)^{-pk/2}$ poses no problem. It is therefore sufficient to show that we have geometric decay in k for the following expression:

$$2^{kv}\left(\frac{1}{(wp/2)+1-wp\epsilon}\right)^{kv/w}\int_1^{e-\delta_k}\left[\frac{e\ln\alpha+\delta_k}{\alpha^{1+(vp/2)}}\right]^k d\alpha.$$

When $p > 2$, arguing as before we need only find v and w satisfying $v^{-1} + w^{-1} = 1$ such that the following expression is bounded by $1 - \eta$ for some $\eta > 0$:

$$\frac{2^v e \ln \alpha}{\alpha^{1+(vp/2)}} \frac{1}{[w((p/2)-1)+1]^{v/w}} \leq \frac{2^v}{(vp/2)+1} \frac{1}{[w((p/2)-1)+1]^{v/w}}.$$

However, as $v \to 1$ and $w \to \infty$ we have

$$\frac{2^v}{(vp/2)+1} \frac{1}{[w((p/2)-1)+1]^{v/w}} \to \frac{4}{p+2},$$

which for $p > 2$ is strictly less than 1. When $p = 2$ and $0 \leq \epsilon < 1/2$ is given, it suffices to find a value of $v > 1$ such that

$$\frac{2^v}{v+1} \frac{1}{\left[\frac{v}{v-1}(1-2\epsilon)+1\right]^{v-1}} < 1. \tag{3.8}$$

When $v = (3/2) - \epsilon$, the expression in (3.8) becomes

$$\frac{2^{(5/2)-\epsilon}}{5-2\epsilon}(4-2\epsilon)^{\epsilon-(1/2)} = \frac{4}{5-2\epsilon}(2-\epsilon)^{-[(1/2)-\epsilon]} =: f(\epsilon).$$

One can easily check that $f(1/2) = 1$ and that f is strictly increasing on $[0, 1/2]$; therefore, together they imply that (3.8) is satisfied when $v = (3/2) - \epsilon$. □

Chapter 4

Consequences of process convergence

Let Γ be the collection of all nonempty paths (finite or infinite) from the root of the infinite binary tree. For any path $\gamma \in \Gamma$, define $|\gamma| \in \{1, 2, \ldots\} \cup \{\infty\}$ to be the number of levels visited by γ, and let $S_{\gamma,n} := \sum_{\theta \in \gamma} S_{\theta,n}$ and $I_\gamma := \sum_{\theta \in \gamma} I_\theta$. Let $p \in [1, \infty)$, and let $a_k \equiv a_{k,p} \equiv 2^{k/p}$ be used in the definition of $|||\cdot|||_p$.

Proposition 4.1. *Let $1 \le p < \infty$ and suppose $\left|\left|\left|\frac{S_n}{n} - I\right|\right|\right|_p \to 0$. Then as $n \to \infty$ we have the L^p-convergence*

$$\left\| \sup_{\gamma \in \Gamma} \left| \frac{S_{\gamma,n}}{n} - I_\gamma \right| \right\|_p \to 0.$$

Proof. Consider a path $\gamma \in \Gamma$ and let $\gamma = (\theta_0, \theta_1, \ldots, \theta_{|\gamma|-1})$ if $|\gamma| < \infty$ and $\gamma =$

$(\theta_0, \theta_1, \ldots)$ if $|\gamma| = \infty$. Then

$$\begin{aligned}
\left|\frac{S_{\gamma,n}}{n} - I_\gamma\right| &= \left|\sum_{0\le k<|\gamma|}\left(\frac{S_{\theta,n_k}}{n} - I_{\theta_k}\right)\right| \\
&= \left|\sum_{0\le k<\infty}\mathbf{1}(k<|\gamma|)\left(\frac{S_{\theta,n_k}}{n} - I_{\theta_k}\right)\right| \\
&\le \sum_{0\le k<\infty}\mathbf{1}(k<|\gamma|)\left|\frac{S_{\theta,n_k}}{n} - I_{\theta_k}\right|, \qquad (4.1)
\end{aligned}$$

where θ_k is defined arbitrarily for $k \ge |\gamma|$ if $|\gamma| < \infty$. Additionally for any k we have

$$\left|\frac{S_{\theta,n_k}}{n} - I_{\theta_k}\right| \le \left(\sum_{\theta\in\Lambda_k}\left|\frac{S_{n,\theta}}{n} - I_\theta\right|^p\right)^{1/p}. \qquad (4.2)$$

Note that the bound in (4.2) does not depend on which path $\gamma \in \Gamma$ was chosen. By combining (4.1)–(4.2), we get the following bound:

$$\begin{aligned}
\left\|\sup_{\gamma\in\Gamma}\left|\frac{S_{\gamma,n}}{n} - I_\gamma\right|\right\|_p &\le \left\|\sum_{0\le k<\infty}\left(\sum_{\theta\in\Lambda_k}\left|\frac{S_{n,\theta}}{n} - I_\theta\right|^p\right)^{1/p}\right\|_p \\
&\le \sum_{0\le k<\infty}\left\|\left(\sum_{\theta\in\Lambda_k}\left|\frac{S_{n,\theta}}{n} - I_\theta\right|^p\right)^{1/p}\right\|_p \\
&= \sum_{0\le k<\infty}\left(\sum_{\theta\in\Lambda_k}\left\|\frac{S_{n,\theta}}{n} - I_\theta\right\|_p^p\right)^{1/p} \\
&\le \sum_{0\le k<\infty} 2^{k/p}\max_{\theta\in\Lambda_k}\left\|\frac{S_{n,\theta}}{n} - I_\theta\right\|_p \\
&= \left|\left|\left|\frac{S_n}{n} - I\right|\right|\right|_p.
\end{aligned}$$

□

Remark 4.2. The same proof shows that Proposition 4.1 continues to hold when Γ is enlarged to include all *random* paths defined on the same probability space as the seeds U_i.

We now focus our attention on the analysis of worst-case `QuickSelect`. Let

$$T_n := n^{-1} \max_{1 \le m \le n} \texttt{FIND}(n, m), \tag{4.3}$$

where $\texttt{FIND}(n, m)$ is the cost for `QuickSelect` to find the mth smallest element of the first n keys. Note that T_n has the following representation:

$$T_n = \sup_{\gamma \in \Gamma} \frac{S_{\gamma,n}}{n} = \sup_{\gamma \in \Gamma} \sum_{\theta \in \gamma} \frac{S_{\theta,n}}{n}. \tag{4.4}$$

Define

$$T := \sup_{\gamma \in \Gamma} I_\gamma. \tag{4.5}$$

Corollary 4.3. *Suppose* $\left\| \left\| \frac{S_n}{n} - I \right\| \right\|_p \to 0$. *Then as* $n \to \infty$ *we have*

$$T_n \xrightarrow{L^p} T.$$

Proof. By Lemma 2.11, we have

$$\|T_n - T\|_p = \left\| \sup_{\gamma \in \Gamma} \frac{S_{\gamma,n}}{n} - \sup_{\gamma \in \Gamma} I_\gamma \right\|_p \le \left\| \sup_{\gamma \in \Gamma} \left| \frac{S_{\gamma,n}}{n} - I_\gamma \right| \right\|_p,$$

and the result follows from Proposition 4.1. □

Corollary 4.4 (L^p-convergence for worst-case `FIND`)**.** *Let* $2 \le p < \infty$*, and suppose that the source-specific cost function* β *is* ϵ*-tame with* $0 \le \epsilon < 1/p$*. Then, recalling* (4.3)–(4.5)*, the scale-normalized cost* T_n *of worst-case* `FIND` *satisfies*

$$T_n \xrightarrow{L^p} T.$$

Proof. Combine Theorem 3.1 and Corollary 4.3. □

4.1 QuickSelect for fixed quantile(s)

4.1.1 QuickVal and QuickQuant

Fix $\alpha \in [0,1]$ and let (m_n) be a sequence of integers satisfying $1 \le m_n \le n$ for every n and $m_n/n \to \alpha$ as $n \to \infty$. Consider the algorithm `QuickQuant`(n, α), defined as `QuickSelect`(n, m_n) applied to the first n keys. Vallée *et al.* [35] introduced, and Fill and Nakama [12] further studied, an algorithm `QuickVal`(n, α) closely related to `QuickQuant` described briefly as follows (see, e.g., [12, Section 2.3] for a more complete description): While `QuickQuant`(n, α) searches successfully for the sample α-quantile among the first n seeds $U_1, \dots, U_n$, the algorithm `QuickVal`(n, α) searches (unsuccessfully, with probability 1) through the seeds $U_1, \dots, U_n$ for the value α.

Let $\gamma(\alpha) \in \Gamma$ be the infinite path from the root seed to seed α in the infinite binary search tree of seeds. By combining Theorem 3.1 and Remark 4.2, we obtain the following result.

Proposition 4.5 (L^p-convergence for `QuickVal`). *Let $2 \le p < \infty$, and suppose that the source-specific cost function β is ϵ-tame with $0 \le \epsilon < 1/p$. Then the cost V_n of* `QuickVal`(n, α) *satisfies*

$$\frac{V_n}{n} \xrightarrow{L^p} I_{\gamma(\alpha)} \quad \text{as } n \to \infty.$$

Remark 4.6. (a) This result is obtained effortlessly from our master Theorem 3.1 but is slightly weaker than Theorem 3.1 in [12] in two ways. First, we require $p \ge 2$, whereas only $p \geq 1$ is assumed in [12]. Second, our tameness hypothesis of $0 \leq \epsilon < 1/p$ is sufficient (but not necessary) for the hypothesis

$$\sum_{k \geq 1} (\mathbb{E} I_{p,\theta_k})^{1/p} < \infty$$

of [12, Theorem 3.1], where $\gamma(\alpha) = (\theta_1, \theta_2, \dots)$ is the random path traversed by `QuickVal`(n, α). (Note the adjustment in indices needed to align notation with [12].) This sufficiency can be proved as follows. As argued at the start of the proof of Lemma 2.4,

$$I_{p,\theta_k} \leq \frac{2^{p\epsilon} c^p}{1 - p\epsilon} (R_{\theta_k} - L_{\theta_k})^{1-p\epsilon}.$$

But according to [12, Lemma 3.5] we have

$$\mathbb{E}(R_{\theta_k} - L_{\theta_k})^{1-p\epsilon} \leq \left(\frac{2 - 2^{-(1-p\epsilon)}}{2 - p\epsilon} \right)^k,$$

and the fraction $[2 - 2^{-(1-p\epsilon)}]/(2 - p\epsilon)$ here is strictly smaller than 1.

(b) Similarly, the next result is Theorem 4.1 in [12], but is proved there assuming only $p \geq 1$.

Theorem 4.7 (L^p-convergence for `QuickQuant`). *Let $2 \leq p < \infty$, and suppose that the source-specific cost function β is ϵ-tame with $0 \leq \epsilon < 1/p$. Then the cost Q_n of* `QuickQuant`(n, α) *satisfies*

$$\frac{Q_n}{n} \xrightarrow{L^p} I_{\gamma(\alpha)} \quad \text{as } n \to \infty.$$

Proof. For each n, let $\gamma_n \equiv \gamma_n(\alpha) = (\theta_{n,0}, \theta_{n,1}, \theta_{n,2}, \dots, \theta_{n,|\gamma_n|})$ be the random path taken by `QuickQuant`(n, α) and let $\gamma \equiv \gamma(\alpha) = (\theta_0, \theta_1, \dots)$ be the random path taken by $\texttt{QuickVal}(n, \alpha)$. It follows from the random-paths extension of Proposition 4.1 mentioned in Remark 4.2 that

$$\left\| \frac{Q_n}{n} - I_{\gamma_n} \right\|_p \to 0.$$

Therefore, by the triangle inequality for L^p-norm, it suffices to show that

$$\|I_{\gamma_n} - I_\gamma\|_p \leq \sum_{k=0}^{\infty} \left\| \mathbf{1}(k < |\gamma_n|) I_{\theta_{n,k}} - I_{\theta_k} \right\|_p \to 0.$$

As an easy consequence of the strong law of large numbers, for each k we have almost surely that

$$\mathbf{1}(k < |\gamma_n|) I_{\theta_{n,k}} - I_{\theta_k} = 0 \text{ for all large } n.$$

Therefore by applying the dominated convergence theorem twice (first for expectation and then for counting measure), we finish the proof. It suffices to show that the upper bound in the inequality

$$\left\| \mathbf{1}(k < |\gamma_n|) I_{\theta_{n,k}} - I_{\theta_k} \right\|_p \leq 2 \left\| \max_{\theta \in \Lambda_k} I_\theta \right\|_p$$

is summable over k. Indeed, by applying Lemmas 2.12 and 2.4 we get the bound

$$\left\| \max_{\theta \in \Lambda_k} I_\theta \right\|_p \leq 2^{k/p} \max_{\theta \in \Lambda_k} \|I_\theta\|_p \leq \left(\frac{c 2^\epsilon}{1-\epsilon} \right) \left(\frac{p+1-p\epsilon}{2} \right)^{-k/p}.$$

Since $\epsilon < 1/p$ and $p \geq 2$, we have that $p + 1 - p\epsilon > 2$, which ensures summability over k. □

4.1.2 MultipleQuickQuant

We will be very brief in this subsection. Consult, for example, [28] for a description of the algorithm `MultipleQuickSelect`$(n; m_1, \ldots, m_t)$ for finding simultaneously the keys of rank $m_1, \ldots, m_t$ in an input sequence of length n.

Now fix a positive integer t and, for $i = 1, \ldots, t$, values $\alpha^i \in [0,1]$ and sequences (m_n^i) of integers satisfying $1 \le m_n^i \le n$ and $m_n^i/n \to \alpha^i$ as $n \to \infty$. Consider the algorithm `MultipleQuickQuant`$(n; \alpha^1, \ldots, \alpha^t)$, defined as `MultipleQuickSelect`$(n; m_n^1, \ldots, m_n^t)$ applied to the first n keys. Let $\gamma \equiv \gamma(\alpha^1, \ldots, \alpha^t)$ be the union of the sets $\gamma(\alpha^i)$ described just before Proposition 4.5, and let $I_\gamma := \sum_{\theta \in \gamma} I_\theta$. The next theorem generalizes Theorem 4.7.

Theorem 4.8 (L^p-convergence for `MultipleQuickQuant`). *Let $2 \le p < \infty$, and suppose that the source-specific cost function β is ϵ-tame with $0 \le \epsilon < 1/p$. Then the cost M_n of* `MultipleQuickQuant`$(n; \alpha^1, \ldots, \alpha^t)$ *satisfies*

$$\frac{M_n}{n} \xrightarrow{L^p} I_\gamma \quad \text{as } n \to \infty.$$

In the interest of brevity, the proof is omitted. We also leave the statement (which by now should be rather obvious) and proof of L^p-convergence for worst-case `MultipleQuickSelect` to the reader.

4.2 Almost-sure convergence for worst-case FIND

Theorem 4.9 (Almost sure convergence for worst-case `FIND`)**.** *If the source-specific cost function β is ϵ-tame with $0 \le \epsilon < 1/4$, then, recalling* (4.3)–(4.5)*, the scale-normalized cost T_n of worst-case* `FIND` *satisfies*

$$T_n \to T \quad \text{a.s. as } n \to \infty.$$

Remark 4.10. Devroye [9] proved this almost sure convergence in the special case $\beta \equiv 1$ of key comparisons.

Proof of Theorem 4.9. For $0 \le \ell < \infty$, let $\Gamma(\ell)$ be the set of 2^ℓ paths from the root to a node at level ℓ. For $0 \le \ell < \infty$, define

$$T(\ell) := \max_{\gamma \in \Gamma(\ell)} I_\gamma, \qquad V(\ell) := \sum_{k > \ell} \max_{\theta \in \Lambda_k} I_\theta$$

$$T_n(\ell) := \max_{\gamma \in \Gamma(\ell)} \frac{S_{\gamma,n}}{n}, \qquad V_n(\ell) := \sum_{k > \ell} \max_{\theta \in \Lambda_k} \frac{S_{\theta,n}}{n}.$$

Then we have the following inequalities for T and T_n:

$$\begin{aligned} T_n(\ell) &\le T_n \le T_n(\ell) + V_n(\ell), \\ T(\ell) &\le T \le T(\ell) + V(\ell), \end{aligned}$$

and hence

$$\begin{aligned} |T_n - T| &\le |T_n - T_n(\ell)| + |T_n(\ell) - T(\ell)| + |T(\ell) - T| \\ &\le V_n(\ell) + |T_n(\ell) - T(\ell)| + V(\ell). \end{aligned}$$

We prove that $T_n \to T$ almost surely in three steps. First, we show for each fixed ℓ that $T_n(\ell) \to T(\ell)$ a.s. as $n \to \infty$, then we show that $V(\ell) \to 0$ a.s. as $\ell \to \infty$, and lastly we show that $\limsup_{\ell\to\infty} \limsup_{n\to\infty} V_n(\ell) = 0$ almost surely. Therefore the result follows from Lemmas 4.11–4.13. □

Lemma 4.11. *We have*

$$T_n(\ell) \to T(l) \quad \text{a.s. as } n \to \infty.$$

Proof. Let $\theta \in \Theta$. Condition on the random vector C_θ; recall that $S_{\theta,n}$ is then the sum of $n - \tau_\theta$ iid random variables and hence

$$\frac{S_{\theta,n}}{n} \to I_\theta \quad \text{a.s.} \tag{4.6}$$

Since (4.6) holds conditionally for (almost) every value of C_θ, it also holds unconditionally, by Tonelli's theorem. Therefore (summing over $\theta \in \gamma$) $S_{n,\gamma}/n \to I_\gamma$ a.s. for each path $\gamma \in \Gamma(\ell)$ because each such path is finite. Then taking the maximum over the finite set of such paths, we conclude the desired result. □

Lemma 4.12. *If β is ϵ-tame with $0 < \epsilon < 1$, then*

$$V(\ell) \to 0 \quad \text{a.s. as } \ell \to \infty.$$

Proof. Let $p \in [1, \infty)$. Using Markov's inequality, for any $\eta > 0$ we have $\mathbb{P}(V(\ell) \geq \eta) \leq \mathbb{E}V(\ell)^p/\eta^p$. Therefore, by the first Borel–Cantelli lemma, it suffices to show that

$$\sum_{\ell=0}^{\infty} \mathbb{E}V(\ell)^p < \infty.$$

However, applying the definition of $V(\ell)$, the triangle inequality for (and continuity of) $\|\cdot\|_p$, the Max Lemma (Lemma 2.12), and Lemma 2.4, we find

$$\|V(\ell)\|_p \le C \sum_{k>\ell} \left(\frac{2}{p+1-p\epsilon}\right)^{k/p},$$

for a constant C not depending on ℓ. This implies that

$$\begin{aligned}\sum_{\ell=0}^{\infty} \mathbb{E}V(\ell)^p &\le C^p \sum_{\ell=0}^{\infty} \left[\left(\frac{2}{p+1-p\epsilon}\right)^{\ell/p} \sum_{k=1}^{\infty} \left(\frac{2}{p+1-p\epsilon}\right)^{k/p}\right]^p \\ &= C^p \left[\sum_{\ell=0}^{\infty} \left(\frac{2}{p+1-p\epsilon}\right)^{\ell}\right] \left[\sum_{k=1}^{\infty} \left(\frac{2}{p+1-p\epsilon}\right)^{k/p}\right]^p.\end{aligned}$$

Choosing p so that $\epsilon < (p-1)/p$, we have $\sum_\ell \mathbb{E}V(\ell)^p < \infty$. □

Lemma 4.13. *If β is ϵ-tame with $0 \le \epsilon < 1/4$, then*

$$\limsup_{\ell\to\infty} \limsup_{n\to\infty} S_n^{(V)}(\ell) = 0 \quad \text{a.s.}$$

Proof. First observe that $S_n^{(V)}(\ell) \le V(\ell) + \widetilde{V}_n(\ell)$, with

$$\widetilde{V}_n(\ell) := \sum_{k>\ell} \max_{\theta\in\Lambda_k} \left[\left|\frac{S_{\theta,n}}{n} - \left(\frac{n-\tau_\theta}{n}\right) I_\theta\right| \mathbf{1}(\tau_\theta < n)\right].$$

In light of Lemma 4.12, it is sufficient to prove that $\limsup_{\ell\to\infty} \limsup_{n\to\infty} \widetilde{V}_n(\ell)$ vanishes almost surely. In fact we will prove the stronger result that $\widetilde{V}_n(\ell) \to 0$ a.s. as $n \to \infty$ for each ℓ. Using the technique of Markov's inequality and first Borel–Cantelli lemma as in Lemma 4.12, it suffices to show that for some $2 \le p < \infty$ we have

$$\sum_{n=1}^{\infty} \mathbb{E}\, \widetilde{V}_n(\ell)^p < \infty.$$

Arguing as in the proof of Lemma 4.12, we have the following bound for any $p \in [2, \infty)$:

$$\mathbb{E}\widetilde{V}_n(\ell)^p \le \left[\sum_{k>\ell} 2^{k/p} \max_{\theta \in \Lambda_k} h(n,p,\theta) \right]^p$$

where $h(n,p,\theta)$ is the first term on the right in inequality (3.3). By the arguments following (3.3) in the proof of Theorem 3.1, we have the following bound on $h(n,p,\theta)$:

$$h(n,p,\theta)^p \le c_p \, \mathbb{E} \left[\mathbf{1}(\tau_\theta < n) \left(\frac{2^p I_{p,\theta}}{n^{p-1}} + \frac{I_{2,\theta}^{p/2}}{n^{p/2}} \right) \right].$$

So we have the following bound:

$$\mathbb{E}\widetilde{V}_n(\ell)^p \le c_p \left[\sum_{k>\ell} \left(2^k \, \mathbb{E} \left[\mathbf{1}(\tau_{\theta_k} < n) \left(\frac{2^p I_{p,\theta_k}}{n^{p-1}} + \frac{I_{2,\theta_k}^{p/2}}{n^{p/2}} \right) \right] \right)^{1/p} \right]^p,$$

where $\theta_k \in \Lambda_k$ is chosen arbitrarily for each k. Choose p such that $\epsilon < 1/p < 1/4$ and then a such that $1 < a < (p/2) - 1$. After factoring out n^{-a}, we are left with the following bound for $\mathbb{E}\widetilde{V}_n(\ell)^p$:

$$\mathbb{E}\widetilde{V}_n(\ell)^p \le c_p n^{-a} \left[\sum_{k>\ell} \left(2^k \, \mathbb{E} \left[\frac{2^p I_{p,\theta_k}}{\tau_{\theta_k}^{p-1-a}} + \frac{I_{2,\theta_k}^{p/2}}{\tau_{\theta_k}^{(p/2)-a}} \right] \right)^{1/p} \right]^p. \qquad (4.7)$$

Note that the only dependence on n in this bound for $\mathbb{E}\widetilde{V}_n(\ell)^p$ is in the factor n^{-a}. Therefore, since $a > 1$ it suffices to prove that, for any ℓ, the sum over k in (4.7) is finite. This can be done by two applications of Lemma 2.9, just as in the proof of Theorem 3.1; the remaining details are routine and omitted. □

Note that the tameness condition $\epsilon < 1/4$ imposed on β for almost sure convergence of worst-case `QuickSelect` in Theorem 4.9 matches the condition imposed for almost sure convergence of `QuickVal` proven by Fill and Nakama in [12, Theorem 3.4] for fixed quantile α.

Chapter 5

Exact L^2 Asymptotics for `QuickMin` Residual

Before deriving a limit law for `QuickVal` under general source-specific cost functions β, we motivate the scaling factor of $\sqrt{n}$ in Theorem 6.4. We consider the case of `QuickMin` (i.e., `QuickSelect` for the minimum key) with key-comparisons cost ($\beta \equiv 1$). Note that the operation of `QuickMin` and `QuickVal` with $\alpha = 0$ are identical. The remainder of this chapter is spent establishing Theorem 5.1, which gives exact and asymptotic expansions for the second moment of the residual in this special case.

Let K_n denote the key-comparisons cost of `QuickMin` and define

$$Y_n := \frac{K_n - \mu_n}{n+1}, \tag{5.1}$$

where $\mu_n := \mathbb{E}K_n = 2(n - H_n)$ for each n [24]. A consequence of [22, Theorem 1] is that $K_n/n \xrightarrow{\mathcal{L}} D$, where $D \stackrel{\mathcal{L}}{=} \sum_{k=0}^{\infty} \prod_{j=0}^{k} U_j$ has a Dickman distribution [22], with $\mu := \mathbb{E}D = 2$ (here $U_0 := 1$). (Note that [22] refers to $D - 1$ as having a Dickman distribution; we ignore this distinction.) Applying [12, Theorems 3.1 and 3.4] to the special case of `QuickMin` using key-comparisons costs yields the stronger result that Y_n converges to a limit random variable Y in L^p for any $p \geq 1$ and almost surely. We can then set

$$D := Y + \mu = Y + 2, \tag{5.2}$$

and this D has a Dickman distribution as defined above.

The main result of this chapter is the exact calculation of of the second moment of $Y_n - Y$:

Theorem 5.1. *For Y_n and Y defined previously, we have*

$$\begin{aligned} a_n^2 := \mathbb{E}(Y_n - Y)^2 &= (n+1)^{-2} \left[\frac{3}{2} n + 4H_n - 4H_n^{(2)} + \frac{1}{2} \right] \\ &= \frac{3}{2} n^{-1} + O\left(\frac{\log n}{n^2} \right). \end{aligned}$$

Define

$$I_n := \#\{1 < i \leq n : U_i < U_1\} \tag{5.3}$$

(i.e., the number of keys that fall into the left subtree of the `QuickSelect` seed process). To begin the derivation, note that $K_n = n - 1 + \tilde{K}_{I_n}$, where $\tilde{K}_{I_n}$ is the key-comparisons cost for `QuickMin` applied to the left subtree. Note also that the same equation holds as equality in law if the process $\tilde{K}$ has the same distribution as the process K and is independent of I_n. We also have $D = 1 + U\tilde{D}$ with $U := U_1$ and $\tilde{D} \stackrel{\mathcal{L}}{=} D$ independent.

Make the following definitions:

$$Y_{n,0} := \frac{K_{I_n} - \mu_{I_n}}{I_n + 1},$$
$$Y^{(0)} := \tilde{D} - \mu. \tag{5.4}$$

Then we can express the residual $Y_n - Y$ in terms of these "smaller versions" of Y_n and Y:

$$\begin{aligned} Y_n - Y &= \frac{n - 1 + K_{I_n} - \mu_n}{n+1} - \left(1 + \ U\tilde{D} - \mu\right) \\ &= \left(\frac{I_n + 1}{n+1}\right) Y_{n,0} - UY^{(0)} + \frac{n-1}{n+1} - 1 + \frac{\mu_{I_n} - \mu_n}{n+1} - U\mu + \mu \\ &= \left(\frac{I_n + 1}{n+1}\right) Y_{n,0} - UY^{(0)} + C_n(I_n)\frac{n}{n+1} - C(U), \end{aligned} \tag{5.5}$$

where $C_n(i) := n^{-1}(n - 1 + \mu_i - \mu_n)$ and $C(x) := \mu x - 1 = 2x - 1$. Observe that with these definitions, we can break up the previous equation as

$$Y_n - Y = W_1 + W_2, \tag{5.6}$$

where

$$W_1 := \frac{I_n + 1}{n+1} Y_{n,0} - UY^{(0)}, \quad W_2 := C_n(I_n)\frac{n}{n+1} - C(U).$$

Conditionally given I_n and U, the random variable W_2 is constant and W_1 has mean zero, so

$$a_n^2 = \mathbb{E}\left(Y_n - Y\right)^2 = \mathbb{E}W_1^2 + \mathbb{E}W_2^2.$$

Consider the first term $\mathbb{E}W_1^2$.

Lemma 5.2.

$$\mathbb{E}W_1^2 = \frac{1}{n}\sum_{k=0}^{n-1}\frac{(k+1)^2}{(n+1)^2}a_k^2 + \frac{1}{12(n+1)}.$$

Proof. If we define

$$Z_1 := \frac{I_n+1}{n+1}\left(Y_{n,0} - Y^{(0)}\right), \quad Z_2 := \left(\frac{I_n+1}{n+1} - U\right)Y^{(0)},$$

then $W_1 = Z_1 + Z_2$ and so $\mathbb{E}W_1^2 = \mathbb{E}Z_1^2 + \mathbb{E}Z_2^2 + 2\mathbb{E}(Z_1Z_2)$.

For the cross term $\mathbb{E}(Z_1Z_2)$, conditionally given I_n the random variable U is distributed $\text{Beta}(I_n+1, n-I_n)$. Therefore,

$$\mathbb{E}(Z_1Z_2) = \mathbb{E}\left\{\mathbb{E}\left[Z_1\left(\frac{I_n+1}{n+1} - U\right)Y^{(0)}\middle| I_n, Y_{n,0}, Y^{(0)}\right]\right\} = 0.$$

Next consider the term $\mathbb{E}Z_1^2$.

Remark 5.3. The conditional joint distribution of the process $(Y_{n,0})_{n\geq 0}$ and the random variable $Y^{(0)}$ given I_n is the conditional joint distribution of the process $(Y^*_{I_n})_{n\geq 0}$ and the random variable Y^* given I_n, where the process (Y^*_n) and the random variable Y^* are independent of I_n and have (unconditionally) the same joint distribution as the process (Y_n) and the random variable Y.

In light of the preceding remark,

$$\mathbb{E}Z_1^2 = \mathbb{E}\left[\left(\frac{I_n+1}{n+1}\right)^2\left(Y^*_{I_n} - Y^*\right)^2\right].$$

Since $I_n \sim \text{unif}\,\{0, 1, 2, \ldots, n-1\}$, conditioning on I_n gives

$$\mathbb{E}Z_1^2 = \frac{1}{n}\sum_{k=0}^{n-1}\left(\frac{k+1}{n+1}\right)^2 a_k^2.$$

Finally, consider the term $\mathbb{E}Z_2^2$. Since $Y^{(0)}$ is independent of I_n and U, we have

$$\mathbb{E}Z_2^2 = \mathbb{E}{Y^{(0)}}^2\, \mathbb{E}\left(\frac{I_n+1}{n+1} - U\right)^2.$$

Recall that $I_n \sim \text{unif}\{0,1,\ldots,n-1\}$ and that conditionally given I_n, we have $U \sim \text{Beta}(I_n+1, n-I_n)$; therefore,

$$\mathbb{E}\left(\frac{I_n+1}{n+1} - U\right)^2 = \frac{1}{n}\sum_{k=0}^{n-1}\frac{(k+1)(n-k)}{(n+1)^2(n+2)} = \frac{1}{6(n+1)}. \tag{5.7}$$

Since $\mathbb{E}{Y^{(0)}}^2 = 1/2$ [23], we have that

$$\mathbb{E}Z_2^2 = \frac{1}{12(n+1)}.$$

Putting these calculations together, we get that

$$\mathbb{E}W_1^2 = \frac{1}{n}\sum_{k=0}^{n-1}\frac{(k+1)^2}{(n+1)^2}a_k^2 + \frac{1}{12(n+1)}.$$

□

We will need the following result in the sequel; it can be proved by reversing the order of summation.

Lemma 5.4. *Let $(a_k)_{k\geq 1}$ be a sequence of real numbers, then*

$$\sum_{m=0}^{n-1}\sum_{k=1}^{m} a_k = \sum_{k=1}^{n-1}(n-k)a_k.$$

Now we consider the term $\mathbb{E}W_2^2$.

Lemma 5.5.

$$\mathbb{E}W_2^2 = \frac{2}{3(n+1)} + \left(\frac{2}{n+1}\right)^2\left(1 - \frac{1}{n}H_n\right).$$

Proof. We have

$$\begin{aligned}
W_2 &= C_n(I_n)\frac{n}{n+1} - C(U) \\
&= \frac{n}{n+1}\left[\frac{1}{n}\left(n-1+2(I_n-H_{I_n})-2(n-H_n)\right)\right]+1-2U \\
&= \frac{1}{n+1}\left[2(I_n-H_{I_n})-n+2H_n-1\right]+1-2U \\
&= 2\left(\frac{I_n+1}{n+1}-U\right)-\frac{2}{n+1}+\frac{1}{n+1}\left(-n+2H_n-2H_{I_n}-1\right)+1 \\
&= 2\left(\frac{I_n+1}{n+1}-U\right)+\frac{2}{n+1}\left(H_n-H_{I_n}-1\right). \qquad (5.8)
\end{aligned}$$

Squaring and then taking expectations, we find

$$\begin{aligned}
\mathbb{E}W_2^2 =& 4\mathbb{E}\left(\frac{I_n+1}{n+1}-U\right)^2+\mathbb{E}\left[\left(\frac{2}{n+1}\right)^2\left(H_n-H_{I_n}-1\right)^2\right] \\
&+4\mathbb{E}\left[\left(\frac{I_n+1}{n+1}-U\right)\left(\frac{2}{n+1}\left(H_n-H_{I_n}-1\right)\right)\right].
\end{aligned}$$

Recall that conditionally given I_n we have $U \sim \text{Beta}(I_n+1, n-I_n)$, which implies that the cross term vanishes; therefore, it suffices to consider the two squared terms. From (5.7) we know that

$$4\mathbb{E}\left(\frac{I_n+1}{n+1}-U\right)^2 = \frac{2}{3(n+1)}.$$

We are now prepared to treat the final term

$$\begin{aligned}
L_n &:= \mathbb{E}\left[\left(\frac{2}{n+1}\right)^2\left(H_n-H_{I_n}-1\right)^2\right] \\
&= \left(\frac{2}{n+1}\right)^2\left[(H_n-1)^2-2(H_n-1)\mathbb{E}H_{I_n}+\mathbb{E}H_{I_n}^2\right]. \qquad (5.9)
\end{aligned}$$

Since $I_n \sim \text{unif}\{0, 1, \ldots, n-1\}$, we have

$$\mathbb{E}H_{I_n} = \frac{1}{n}\sum_{m=0}^{n-1}\sum_{k=1}^{m}\frac{1}{k} = \frac{1}{n}\sum_{k=1}^{n-1}\frac{n-k}{k} = H_{n-1} - \frac{n-1}{n}, \tag{5.10}$$

where the second equality follows from Lemma 5.4. It remains to consider

$$\begin{aligned}
\mathbb{E}H_{I_n}^2 &= \sum_{m=0}^{n-1}\frac{1}{n}\left(\sum_{k=1}^{m}\frac{1}{k}\right)^2 \\
&= \frac{1}{n}\sum_{m=0}^{n-1}\left[-\sum_{k=1}^{m}\frac{1}{k^2} + 2\sum_{j=1}^{m}\sum_{i=1}^{j}\frac{1}{ij}\right] \\
&= \frac{1}{n}\left[-\sum_{j=1}^{n-1}\frac{n-j}{j^2} + 2\sum_{m=0}^{n-1}\sum_{j=1}^{m}\sum_{i=1}^{j}\frac{1}{ij}\right] \\
&= -H_{n-1}^{(2)} + \frac{1}{n}H_{n-1} + \frac{2}{n}\sum_{m=0}^{n-1}\sum_{j=1}^{m}\sum_{i=1}^{j}\frac{1}{ij},
\end{aligned}$$

where the penultimate equality follows from Lemma 5.4. Again appealing to Lemma 5.4 gives

$$\begin{aligned}
\frac{1}{n}\sum_{m=0}^{n-1}\sum_{j=1}^{m}\sum_{i=1}^{j}\frac{1}{ij} &= \frac{1}{n}\sum_{m=0}^{n-1}\sum_{j=1}^{m}\frac{1}{j}H_j \\
&= \frac{1}{n}\sum_{k=1}^{n-1}\frac{n-k}{k}H_k \\
&= \sum_{k=1}^{n-1}\left(\frac{1}{k}H_k - \frac{1}{n}H_k\right) \\
&= \sum_{k=1}^{n-1}\frac{1}{k}H_k - \frac{1}{n}\sum_{k=1}^{n-1}\sum_{j=1}^{k}\frac{1}{j} \\
&= \sum_{k=1}^{n-1}\frac{1}{k}H_k - \left(H_{n-1} - \frac{n-1}{n}\right).
\end{aligned}$$

Plugging the last two results and (5.10) into (5.9) gives

$$L_n = \left(\frac{2}{n+1}\right)^2 \left[(H_n - 1)^2 - 2(H_n - 1)\left(H_{n-1} - \frac{n-1}{n}\right)\right.$$
$$\left. -H^{(2)}_{n-1} + \frac{1}{n}H_{n-1} - 2\left(H_{n-1} - \frac{n-1}{n} - \sum_{k=1}^{n-1}\frac{1}{k}H_k\right)\right].$$

Note that

$$H^{(2)}_{n-1} = \sum_{k=1}^{n-1}\frac{1}{k^2},$$

so that

$$-H^{(2)}_{n-1} + 2\sum_{k=1}^{n-1}\sum_{j=1}^{k}\frac{1}{kj} = \left(\sum_{k=1}^{n-1}\frac{1}{k}\right)^2.$$

This allows us to simplify the previous expression:

$$L_n = \left(\frac{2}{n+1}\right)^2 \left[(H_n - 1)^2 - 2(H_n - 1)\left(H_{n-1} - \frac{n-1}{n}\right)\right.$$
$$\left. +H^2_{n-1} + \frac{1}{n}H_{n-1} - 2\left(H_{n-1} - \frac{n-1}{n}\right)\right].$$

After applying some algebraic manipulation, we can simplify the above expression:

$$\begin{aligned}
\left(\frac{n+1}{2}\right)^2 L_n &= H_n^2 - 2H_n + 1 - 2\left[H_n H_{n-1} - H_{n-1} + \left(\frac{1}{n} - 1\right) H_n - \left(\frac{1}{n} - 1\right)\right] \\
&\qquad + H_{n-1}^2 + \frac{1}{n}H_{n-1} - 2H_{n-1} - 2\left(\frac{1}{n} - 1\right) \\
&= \left(H_n^2 - 2H_n H_{n-1} + H_{n-1}^2\right) - 2H_n + 2H_{n-1} - 2\left(\frac{1}{n} - 1\right) H_n \\
&\qquad + \frac{1}{n}H_{n-1} - 2H_{n-1} + 1 + 2\left(\frac{1}{n} - 1\right) - 2\left(\frac{1}{n} - 1\right) \\
&= (H_n - H_{n-1})^2 - 2H_n + 2\left(H_n - \frac{1}{n}\right) - 2\left(\frac{1}{n} - 1\right) H_n \\
&\qquad + \frac{1}{n}\left(H_n - \frac{1}{n}\right) - 2\left(H_n - \frac{1}{n}\right) + 1 \\
&= \frac{1}{n^2} - \frac{2}{n} - 2\left(\frac{1}{n} - 1\right) H_n + \frac{1}{n}H_n - \frac{1}{n^2} - 2H_n + \frac{2}{n} + 1 \\
&= -\frac{2}{n}H_n + \frac{1}{n}H_n + 2H_n - 2H_n + 1 \\
&= -\frac{1}{n}H_n + 1.
\end{aligned}$$

Therefore we get that

$$\mathbb{E}W_2^2 = \frac{2}{3(n+1)} + \left(\frac{2}{n+1}\right)^2 \left(1 - \frac{1}{n}H_n\right),$$

as desired. □

Proof of Theorem 5.1. Combining the expressions for $\mathbb{E}W_1^2$ and $\mathbb{E}W_2^2$ gives

$$(n+1)^2 a_n^2 = \frac{1}{n}\sum_{k=0}^{n-1}(k+1)^2 a_k^2 + \frac{n+1}{12} + \frac{2(n+1)}{3} + 4\left(1 - \frac{1}{n}H_n\right). \tag{5.11}$$

If we define

$$b_n := \frac{n+1}{12} + \frac{2(n+1)}{3} + 4\left(1 - \frac{1}{n}H_n\right) = \frac{3(n+1)}{4} + 4\left(1 - \frac{1}{n}H_n\right),$$

then Lemma 2.15 implies

$$(n+1)^2 a_n^2 = a_0^2 + \sum_{j=1}^{n-1} \frac{b_j}{j+1} + b_n.$$

Plugging in b_n gives

$$(n+1)^2 a_n^2 = a_0^2 + \sum_{j=1}^{n-1} \left[\frac{3(j+1)}{4(j+1)} + \frac{4}{j+1} \left(1 - \frac{1}{j} H_j \right) \right] + \frac{3(n+1)}{4} + 4 \left(1 - \frac{1}{n} H_n \right).$$

Simplifying this expression gives

$$\begin{aligned}
(n+1)^2 a_n^2 &= a_0^2 + \frac{3}{4}(n-1) + \sum_{j=1}^{n-1} \frac{4}{j+1} - \sum_{j=1}^{n-1} \frac{4}{j(j+1)} H_j + \frac{19}{4} \\
&\quad + \frac{3n}{4} - \frac{4}{n} H_n \\
&= \frac{1}{2} + \frac{3}{2} n + 4H_n - \frac{4}{n} H_n - \sum_{j=1}^{n-1} \frac{4}{j(j+1)} H_j \\
&= \frac{1}{2} + \frac{3}{2} n + 4H_n - \frac{4}{n} H_n - 4 \left(H_n^{(2)} - \frac{H_n}{n} \right) \\
&= \frac{3}{2} n + 4H_n - 4H_n^{(2)} + \frac{1}{2},
\end{aligned}$$

where $a_0^2 = 1/2$ was substituted in the second equality. Therefore, we can conclude that

$$\begin{aligned}
a_n^2 = \mathbb{E}(Y_n - Y)^2 &= (n+1)^{-2} \left[\frac{3}{2} n + 4H_n - 4H_n^{(2)} + \frac{1}{2} \right] \\
&= \frac{3}{2} n^{-1} + O\left(\frac{\log n}{n^2} \right).
\end{aligned}$$

□

Chapter 6

Convergence of `QuickVal` Residual

6.1 Preliminaries

The main result in this chapter is Theorem 6.4, which says that the scaled cost of the `QuickVal` residual converges in law to a scale mixture of centered Gaussians. For the remainder of this section, we introduce necessary notation and prove Lemma 6.1, which gives an explicit representation of the mixing distribution. In Section 6.2, we state and prove Theorem 6.4.

Consider a binary search tree (BST) constructed by the insertion (in order) of the n seeds. Then `QuickQuant`(n, α) follows the path from the root to the node storing the m_n^{th} smallest key, where $m_n/n \to \alpha$.

For `QuickVal`(n, α), consider the same BST of seeds with the additional value α inserted (last). Then `QuickVal`(n, α) follows the path from the root to this α-node.

Almost surely for n large and k fixed, the difference between these two algorithms in costs associated with the k-th pivot is negligible to lead order [12, (4.2)]. See [35] or [12] for a more complete description.

Recall definitions (2.1)–(2.6) from Section 2.2, which for the reader's convenience we repeat here:

$$\begin{aligned}
\tau_k &:= \inf\{i :\ L_{k-1} < U_i < R_{k-1}\}, \\
L_k &:= \mathbf{1}(U_{\tau_k} < \alpha)U_{\tau_k} + \mathbf{1}(U_{\tau_k} > \alpha)L_{k-1}, \\
R_k &:= \mathbf{1}(U_{\tau_k} < \alpha)R_{k-1} + \mathbf{1}(U_{\tau_k} > \alpha)U_{\tau_k}, \\
C_k &:= (L_{k-1}, R_{k-1}, \tau_k, U_{\tau_k}), \\
X_{k,i} &:= \mathbf{1}(L_{k-1} < U_i < R_{k-1})\beta(U_i, U_{\tau_k}), \\
S_{k,n} &:= \sum_{i:\, \tau_k < i \le n} X_{k,i}.
\end{aligned}$$

The cost of `QuickVal` on n keys is then given by

$$S_n^{(\mathrm{V})} := \sum_{k=1}^{\infty} S_{k,n}. \tag{6.1}$$

Define

$$\widehat{C}_K := \{C_k : k = 1, \ldots, K\},$$

and

$$\widehat{X}_{K,i} := \sum_{k=1}^{K} X_{k,i}.$$

Then, conditionally given $\widehat{C}_K$, the random variable

$$\widehat{S}_{K,n} := \sum_{\tau_K < i \le n} \widehat{X}_{K,i}$$

is the sum of $(n - \tau_K)^+$ independent and identically distributed random variables, each with the same conditional distribution as $\widehat{X}_K := \sum_{k=1}^{K} X_k$, where

$$X_k := \mathbf{1}(L_{k-1} < U < R_{k-1})\beta(U, U_{\tau_k})$$

and U is uniformly distributed on $(0,1)$ and independent of all the U_j's. Here, $\widehat{X}_{K,i}$ is the cost incurred by comparing seed i to pivots $1, 2, \ldots K$ and $\widehat{S}_{K,n}$ is the comparison cost of all seeds that arrive after the K-th pivot to pivots $1, 2, \ldots K$.

It will be helpful to condition on $\widehat{C}_K$ later. In anticipation of this, we establish notation for the conditional expectation of X_k given C_k (which equals the conditional expectation given $\widehat{C}_k$) and, for $k \le \ell$, the conditional expected product of X_k and X_ℓ given $\widehat{C}_\ell$, as follows:

$$I_k := \mathbb{E}[X_k | C_k] = \int_{L_{k-1}}^{R_{k-1}} \beta(u, U_{\tau_k})\, du, \tag{6.2}$$

$$I_{2,k,\ell} := \mathbb{E}[X_k X_\ell | \widehat{C}_\ell] = \int_{L_{\ell-1}}^{R_{\ell-1}} \beta(u, U_{\tau_k})\beta(u, U_{\tau_\ell})\, du. \tag{6.3}$$

We symmetrize the definition of $I_{2,k,\ell}$ in the indices k and ℓ by setting $I_{2,k,\ell} := I_{2,\ell,k}$ for $k > \ell$. Finally, we write $I_{2,k}$ as shorthand for $I_{2,k,k}$.

We now calculate the mean and variance of $\widehat{X}_K$ with the intention of applying the classical central limit theorem; everything is done conditionally given $\widehat{C}_K$. Define μ_K and σ_K^2 to be the conditional mean and conditional variance of $\widehat{X}_K$ given $\widehat{C}_K$, respectively. Then

$$\mu_K = \sum_{k=1}^{K} I_k, \qquad \sigma_K^2 = \sum_{1 \le k, \ell \le K} \left(I_{2,k,\ell} - I_k I_\ell\right).$$

We next present a condition guaranteeing that σ_K^2 behaves well as $K \to \infty$. We note in passing that this condition is also the sufficient condition of Theorem 3.1 in [12] ensuring that S_n/n converges in L^2 to

$$S := \sum_{k \geq 1} I_k. \tag{6.4}$$

Lemma 6.1. *If*

$$\sum_{k=1}^{\infty} (\mathbb{E} I_{2,k})^{1/2} < \infty, \tag{6.5}$$

then both almost surely and in L^1 we have that (i) *the two series on the right in the equation*

$$\sigma_\infty^2 := \sum_{k=1}^{\infty} (I_{2,k} - I_k^2) + 2 \sum_{\ell=1}^{\infty} \sum_{k=1}^{\ell-1} (I_{2,k,\ell} - I_k I_\ell). \tag{6.6}$$

converge absolutely, (ii) *the equation holds, and* (iii) $\sigma_K \xrightarrow{L^1} \sigma_\infty$ *as* $K \to \infty$.

Proof. Recall the notation $X_k = \mathbf{1}(L_{k-1} < U < R_{k-1}) \beta(U, U_{\tau_k})$ from above. Consider $1 \leq k \leq \ell$. The term $I_{2,k,\ell} - I_k I_\ell$ equals the conditional covariance of X_k and X_ℓ given $\widehat{C}_\ell$, and the absolute value of this conditional covariance is bounded above by the product of the conditional L^2-norms, namely, $I_{2,k}^{1/2} I_{2,\ell}^{1/2}$. Thus for the three desired conclusions it is sufficient that $\mathbb{E}\left(\sum_{k=1}^{\infty} I_{2,k}^{1/2}\right)^2 < \infty$. But

$$\mathbb{E}\left(\sum_{k=1}^{\infty} I_{2,k}^{1/2}\right)^2 = \left\|\sum_{k=1}^{\infty} I_{2,k}^{1/2}\right\|_2^2 \leq \left(\sum_{k=1}^{\infty} \left\|I_{2,k}^{1/2}\right\|_2\right)^2 = \left(\sum_{k=1}^{\infty} (\mathbb{E} I_{2,k})^{1/2}\right)^2.$$

∎

Remark 6.2. In light of the absolute convergence noted in conclusion (i) of Lemma 6.1, we may unambiguously write

$$\sigma_\infty^2 = \sum_{1 \le k, \ell < \infty} (I_{2,k,\ell} - I_k I_\ell), \tag{6.7}$$

both in L^1 and almost surely.

Remark 6.3. Note that if the source-specific cost function β is ϵ-tame for some $\epsilon < 1/2$, then, by Lemma 2.4 with $s = 2$ and $r = 1$, condition (6.5) in Lemma 6.1 is satisfied, because the series there enjoys geometric convergence.

6.2 Convergence

Our main result is that, for a suitably tame cost function, the `QuickVal` residual converges in law to a scale-mixture of centered Gaussians. Furthermore, we have the explicit representation of Lemma 6.1 for the random scale σ_∞ as an infinite series of random variables that depend on conditional variances and covariances related to the source-specific cost functions [see (6.7) and (6.2)–(6.3)].

Theorem 6.4. *Suppose that the cost function β is ϵ-tame with $\epsilon < 1/2$. Then*

$$\sqrt{n}\left(\frac{S_n^{(\mathrm{V})}}{n} - S\right) \xrightarrow{\mathcal{L}} \sigma_\infty Z,$$

where Z has a standard normal distribution and is independent of σ_∞.

We approach the proof of Theorem 6.4 in two parts. First in Proposition 6.7 we apply the central limit theorem to an approximation $\widehat{S}_{K,n}$ of the cost of `QuickVal`

$S_n^{(\mathrm{V})}$. Second we show that the error due to the approximation $\widehat{S}_{K,n}$ is negligible in the limit, culminating in the results of Propositions 6.11 and 6.13.

Before proving Theorem 6.4, we state a corollary to Theorem 6.4 for `QuickMin`. Recall that `QuickMin` is `QuickSelect` applied to find the minimum of the keys. Using a general source-specific cost function β, denote the cost of `QuickMin` on n keys by V_n. Since the operation of `QuickMin` is the same as that of `QuickVal` with $\alpha = 0$, Theorem 6.4 implies the following convergence for the cost of `QuickMin` with the same mild tameness condition on the source-specific cost function.

Corollary 6.5. *Suppose that the source-specific cost function β is ϵ-tame with $\epsilon < 1/2$. Then*

$$\sqrt{n}\left(\frac{V_n}{n} - S\right) \xrightarrow{\mathcal{L}} \sigma_\infty Z,$$

where Z has a standard normal distribution and is independent of σ_∞.

Remark 6.6. In the key-comparisons case $\beta = 1$ (which is ϵ-tame for every $\epsilon \geq 0$) for $k \geq 0$ we have $L_k \equiv 0$ and $R_k \equiv U_{\tau_k}$, with the convention $U_{\tau_0} := 1$. Hence $I_k = U_{\tau_{k-1}}$ for $k \geq 1$, and $I_{2,k,\ell} = U_{\tau_{\ell-1}}$ for $1 \leq k \leq \ell$. Therefore $S = \sum_{k \geq 1} U_{\tau_{k-1}} = 1 + \sum_{k \geq 1} U_{\tau_k}$ and

$$\sigma_\infty^2 = \sum_{1 \leq k, \ell < \infty} (1 - U_{\tau_k}) U_{\tau_\ell}$$

in Corollary 6.5. To further simplify the understanding of σ_∞^2, and hence of the limit in Corollary 6.5 in this case, observe that $U_{\tau_1}, U_{\tau_2}, \ldots$ have the same joint distribution as the cumulative products $U_1, U_1U_2, \ldots$. Thus

$$\sigma_\infty^2 \stackrel{\mathcal{L}}{=} \sum_{1 \le k, \ell < \infty} \left[\left(1 - \prod_{i=1}^{k} U_i \right) \prod_{j=1}^{\ell} U_j \right].$$

Define

$$T_{K,n} := \frac{\widehat{S}_{K,n} - (n - \tau_K)^+ \mu_K}{\sqrt{n}}.$$

Proposition 6.7. *Fix $K \in \{1, 2, \ldots\}$. Suppose that*

$$\mathbb{E} I_{2,k} < \infty$$

for $k = 1, 2, \ldots, K$. Then

$$T_{n,K} \stackrel{\mathcal{L}}{\to} \sigma_K Z$$

as $n \to \infty$, where Z has a standard normal distribution independent of σ_K.

Proof. The classical central limit theorem for independent and identically distributed random variables

applied conditionally given $\widehat{C}_K$ yields

$$\mathcal{L}\left(\frac{\widehat{S}_{K,n} - (n - \tau_K)^+ \mu_K}{\sqrt{(n - \tau_K)^+}} \middle| \widehat{C}_K \right) \stackrel{\mathcal{L}}{\to} \mathrm{N}(0, \sigma_{\mathrm{K}}^2). \tag{6.8}$$

Since τ_K is finite almost surely, Slutsky's theorem (applied conditionally given $\widehat{C}_K$) implies that we can replace $\sqrt{(n - \tau_K)^+}$ by $\sqrt{n}$ in the denominator of (6.8). Finally, applying the dominated convergence theorem to conditional distribution functions gives that the resulting conditional convergence in distribution in (6.8) holds unconditionally. □

Define

$$W_{K,n} := \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{\tau_k < i \leq n} (X_{k,i} - I_k), \qquad \overline{W}_{K,n} := \frac{1}{\sqrt{n}} \sum_{k=K+1}^{\infty} \sum_{\tau_k < i \leq n} (X_{k,i} - I_k),$$

and let

$$W_n := W_{K,n} + \overline{W}_{K,n}.$$

Note that W_n does not depend on K. We can write W_n in terms of the cost of `QuickVal` as follows:

$$W_n = \frac{1}{\sqrt{n}} \left(S_n^{(\mathrm{V})} - \sum_{k=1}^{\infty} (n - \tau_k)^+ I_k \right).$$

We prove that $W_n \xrightarrow{\mathcal{L}} \sigma_\infty Z$ (which is Proposition 6.11) in three parts. First (Lemma 6.8) we show that $|T_{K,n} - W_{K,n}| \to 0$ almost surely. Next (Lemma 6.9) we show that $\left\| \overline{W}_{K,n} \right\|_2$ is negligible as first $n \to \infty$ and then $K \to \infty$. Lastly (see the proof below of Proposition 6.11), an application of Markov's inequality gives the desired convergence.

Lemma 6.8. *For K fixed, if $\mathbb{E} I_k < \infty$ for $k = 1, 2, \ldots, K$, then*

$$|T_{K,n} - W_{K,n}| \to 0$$

almost surely as $n \to \infty$.

Remark The condition $\mathbb{E} I_k < \infty$ in Lemma 6.8 is weaker than the condition $\mathbb{E} I_{2,k} < \infty$ in Proposition 6.7.

Proof. When $n > \tau_K$ we have

$$|T_{K,n} - W_{K,n}| = \frac{1}{\sqrt{n}} \left| \sum_{k=1}^{K} \sum_{\tau_k < i \leq \tau_K} (X_{k,i} - I_k) \right|$$
$$\leq \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{\tau_k < i \leq \tau_K} |X_{k,i} - I_k| .$$

For a fixed K with $k \leq K$, the almost sure finiteness of τ_k and τ_K implies that the sum

$$\sum_{k=1}^{K} \sum_{\tau_k < i \leq \tau_K} |X_{k,i} - I_k| , \tag{6.9}$$

consists of an almost surely finite number of terms. Since each term $|X_{k,i} - I_k|$ is finite almost surely, the sum in (6.9) is finite almost surely. Therefore, $|T_{K,n} - W_{K,n}| \to 0$ almost surely as $n \to \infty$. □

Lemma 6.9. *Let*

$$\epsilon_K := \sum_{k=K+1}^{\infty} (\mathbb{E} I_{2,k})^{1/2} .$$

Then

$$\left\| \overline{W}_{K,n} \right\|_2 \leq \epsilon_K .$$

Remark 6.10. A necessary and sufficient condition for $\epsilon_K \to 0$ as $K \to \infty$ is (6.5). Therefore by Remark 6.3, ϵ-tameness for some $\epsilon < 1/2$ is sufficient.

Proof. Minkowski's inequality yields

$$\left\| \overline{W}_{K,n} \right\|_2 \leq \frac{1}{\sqrt{n}} \sum_{k=K+1}^{\infty} \left\| \sum_{\tau_k < i \leq n} (X_{k,i} - I_k) \right\|_2 . \tag{6.10}$$

By conditioning on C_k, we can calculate the square of the L^2-norm here:

$$\begin{aligned}\Big\| \sum_{\tau_k < i \le n} (X_{k,i} - I_k) \Big\|_2^2 &= \mathbb{E}\,\mathbb{E}\Big[\Big(\sum_{\tau_k < i \le n} (X_{k,i} - I_k)\Big)^2 \Big| C_k\Big] \\ &= \mathbb{E}\left\{(n-\tau_k)^+ (I_{2,k} - I_k^2)\right\} \\ &\le n\mathbb{E} I_{2,k}, \qquad (6.11)\end{aligned}$$

where we use the fact that, conditionally given C_k, the random variables $X_{k,i} - I_k$ for $i > \tau_k$ are iid with zero mean. Substituting (6.11) into (6.10) gives the result. □

Proposition 6.11. *Suppose that*

$$\sum_{k=1}^{\infty} (\mathbb{E} I_{2,k})^{1/2} < \infty.$$

Then

$$W_n \xrightarrow{\mathcal{L}} \sigma_\infty Z,$$

where Z has a standard normal distribution independent of σ_∞.

Proof. Let $t \in \mathbb{R}$ and $\delta > 0$. Since $W_n \le t$ implies either

$$W_{K,n} \le t + \delta \quad \text{or} \quad |W_n - W_{K,n}| > \delta,$$

we have

$$\mathbb{P}[W_n \le t] \le \mathbb{P}[W_{K,n} \le t + \delta] + \mathbb{P}[|W_n - W_{K,n}| > \delta]. \qquad (6.12)$$

Markov's inequality and Lemma 6.9 imply

$$\mathbb{P}[|W_n - W_{K,n}| > \delta] \le \frac{\epsilon_K^2}{\delta^2}. \qquad (6.13)$$

Taking limits superior as $n \to \infty$ gives

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}[W_n \le t] &\le \limsup_{n\to\infty} \mathbb{P}[W_{n,K} \le t+\delta] + \frac{\epsilon_K^2}{\delta^2} \\
&\le \limsup_{n\to\infty} \mathbb{P}[T_{n,K} \le t+2\delta] + \frac{\epsilon_K^2}{\delta^2} \\
&= \mathbb{P}[\sigma_K Z \le t+2\delta] + \frac{\epsilon_K^2}{\delta^2},
\end{aligned}$$

by (6.12)–(6.13), Lemma 6.8, and Proposition 2.4, respectively. Now taking limits as $K \to \infty$ gives

$$\limsup_{n\to\infty} \mathbb{P}[W_n \le t] \le \mathbb{P}[\sigma_\infty Z \le t+2\delta]$$

by Lemma 6.1 and the assumption that $\epsilon_K \to 0$. Letting $\delta \to 0$ yields

$$\limsup_{n\to\infty} \mathbb{P}[W_n \le t] \le \mathbb{P}[\sigma_\infty Z \le t]. \tag{6.14}$$

Applying the previous argument with limsup replaced by liminf to

$$\mathbb{P}[W_n \le t] \ge \mathbb{P}[W_{K,n} \le t-\delta] - \mathbb{P}[|W_n - W_{K,n}| \ge \delta]$$

implies

$$\liminf_{n\to\infty} \mathbb{P}[W_n \le t] \ge \mathbb{P}[\sigma_\infty Z < t]. \tag{6.15}$$

Since $\sigma_\infty Z$ has a continuous distribution, combining (6.14) and (6.15) gives the result.

□

For completeness we include the following simple lemma, which will be needed in the sequel.

Lemma 6.12. *Let $0 < p < 1$ and $a_1, \ldots, a_n$ be nonnegative real numbers. Then*

$$\left(\sum_{k=1}^{n} a_k\right)^p \le \sum_{k=1}^{n} a_k^p.$$

The final step in the proof of Theorem 6.4 is to show that the difference between the centering random variable

$$\sum_{k=1}^{\infty} (n - \tau_k)^+ I_k$$

in W_n and the more natural

$$nS = \sum_{k=1}^{\infty} nI_k$$

is negligible (when scaled by $1/\sqrt{n}$) in the limit as $n \to \infty$.

Proposition 6.13. *If the source-specific cost function β is ϵ-tame with $\epsilon < 1/2$, then*

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{\infty} \left[n - (n - \tau_k)^+\right] I_k \to 0$$

almost surely as $n \to \infty$.

Proof. Observe that for any $0 < \delta < 1/2$, we have

$$[n - (n - \tau_k)^+] = \min(n, \tau_k) \le \tau_k^{(1/2)+\delta} n^{(1/2)-\delta}.$$

Therefore, if we let $0 < \delta < (1/2) - \epsilon$, it suffices to show that

$$\sum_{k=1}^{\infty} \tau_k^{(1/2)+\delta} I_k < \infty \tag{6.16}$$

almost surely. We prove this by showing that the random variable in (6.16) has finite expectation. Applying [12, Lemma 3.6] implies that for the ϵ-tameness constant c, we have

$$I_k \le \frac{2^\epsilon c}{1-\epsilon}(R_{k-1} - L_{k-1})^{1-\epsilon}.$$

Define, for $k = 1, 2, \ldots$, the sigma-field $\mathcal{F}_k := \sigma\langle (L_1, R_1), \ldots (L_{k-1}, R_{k-1})\rangle$. Conditionally given $\mathcal{F}_k$, the distribution of τ_k is the convolution over $j = 0, \ldots, k-1$ of geometric distributions with success probabilities $R_j - L_j$. This distribution is stochastically smaller than the convolution of k geometric distributions with success probability $R_{k-1} - L_{k-1}$. Let $G_k, G_{k,0}, \ldots, G_{k,k-1}$ be $k+1$ iid geometric random variables with success probability $R_{k-1} - L_{k-1}$. Then

$$\begin{aligned}
\mathbb{E}\left[\tau_k^{(1/2)+\delta} I_k \middle| \mathcal{F}_k\right] &\le C_1 \mathbb{E}\left[\left(\sum_{i=0}^{k-1} G_{k,i}\right)^{(1/2)+\delta} (R_{k-1} - L_{k-1})^{1-\epsilon} \middle| L_{k-1}, R_{k-1}\right] \\
&\le C_1 (R_{k-1} - L_{k-1})^{1-\epsilon}\, \mathbb{E}\left[\sum_{i=0}^{k-1} G_{k,i}^{(1/2)+\delta} \middle| L_{k-1}, R_{k-1}\right] \\
&\le C_1 k (R_{k-1} - L_{k-1})^{1-\epsilon}\, \mathbb{E}\left[G_k^{(1/2)+\delta} \middle| L_{k-1}, R_{k-1}\right], \qquad (6.17)
\end{aligned}$$

where

$$C_1 := \frac{2^\epsilon c}{1-\epsilon}.$$

We can now compute

$$\mathbb{E}\left[G_k^{(1/2)+\delta} \middle| L_{k-1}, R_{k-1}\right] = \sum_{i=1}^{\infty} z^{i-1}(1-z) i^p, \qquad (6.18)$$

where $z = 1 - (R_{k-1} - L_{k-1}) \in [0, 1)$ for $k \ge 2$ is the failure probability and $p = (1/2) + \delta$. Note that the infinite series in (6.18) can be written in terms of a polylogarithm function, as follows:

$$\sum_{i=1}^{\infty} z^{i-1}(1-z)i^p = z^{-1}(1-z)\mathrm{Li}_{-p,0}(z), \qquad \mathrm{Li}_{\alpha,r}(z) := \sum_{n=1}^{\infty} (\log i)^r \frac{z^i}{i^\alpha}.$$

Therefore [15, Theorem 1] implies the existence of an $\eta \in (0,1)$ such that for $1-\eta < z < 1$, we have

$$\sum_{i=1}^{\infty} z^i i^p \le \Gamma(1+p)(1-z)^{-(1+p)}.$$

On $0 \le z \le 1-\eta$, the polylogarithm $\mathrm{Li}_{-p,0}(z)$ is increasing and therefore we have the bound

$$\mathrm{Li}_{-p,0}(z) \le \sum_{i=1}^{\infty} (1-\eta)^i i^p =: C_{p,\eta}$$

Defining

$$C_2 := \max(\Gamma(1+p), C_{p,\eta}),$$

for $z \in [0,1)$ we get

$$\mathrm{Li}_{-p,0}(z) \le C_2(1-z)^{-(1+p)}. \tag{6.19}$$

Substituting the bound from (6.19) in (6.18) gives

$$\begin{aligned}\mathbb{E}\left[G_k^p | L_{k-1}, R_{k-1}\right] &\le C_2 \frac{R_{k-1}-L_{k-1}}{1-(R_{k-1}-L_{k-1})}(R_{k-1}-L_{k-1})^{-(1+p)} \\ &= C_2 \sum_{j=0}^{\infty} (R_{k-1}-L_{k-1})^{j-p}.\end{aligned}$$

Therefore, after substituting $p = (1/2)+\delta$, an application of the monotone convergence theorem yields

$$\mathbb{E}(\tau_k^{(1/2)+\delta} I_k) \le C_3 k \sum_{j=0}^{\infty} \mathbb{E}(R_{k-1}-L_{k-1})^{j+(1/2)-\epsilon-\delta},$$

where $C_3 := C_1 C_2$. Let $q := 1/2 - \epsilon - \delta$; then by the restriction placed on δ, we know $q > 0$. By [12, Lemma 3.5], we have

$$\mathbb{E}(\tau_k^{(1/2)+\delta} I_k) \leq C_3 k \sum_{j=0}^{\infty} \left(\frac{2 - 2^{-(j+q)}}{j+q+1} \right)^{k-1}.$$

Therefore, after defining

$$\gamma_j := \frac{2 - 2^{-(j+q)}}{j+q+1},$$

we have

$$\begin{aligned} \sum_{k=3}^{\infty} \mathbb{E}(\tau_k^{(1/2)+\delta} I_k) &\leq C_3 \sum_{k=3}^{\infty} k \sum_{j=0}^{\infty} \gamma_j^{k-1} \\ &= C_3 \sum_{j=0}^{\infty} \sum_{k=3}^{\infty} k \gamma_j^{k-1} \\ &\leq 3 C_3 \sum_{j=0}^{\infty} \frac{\gamma_j^2}{(1-\gamma_j)^2}. \end{aligned}$$

Consequently, to check the convergence in (6.16), it suffices to check that

$$\sum_{j=0}^{\infty} \gamma_j^2 < \infty;$$

however, this follows trivially from the observation that

$$\gamma_j^2 \leq \frac{4}{j^2}.$$

Therefore, it remains to show that the $k = 1, 2$ terms in (6.16) have finite expectation. The first arrival time τ_1 equals 1 identically and $\mathbb{E} I_1 < \infty$. Applying (6.17) when $k = 2$ gives

$$\mathbb{E}\left[\tau_2^{(1/2)+\delta} I_2 \middle| \mathcal{F}_2 \right] \leq 2 C_1 (R_1 - L_1)^{1-\epsilon} \, \mathbb{E}\left[G_2^{(1/2)+\delta} \middle| L_1, R_1 \right].$$

Since $(R_1 - L_1)^{1-\epsilon} < 1$ a.s. , it suffices to show that

$$\mathbb{E}G_2^p < \infty \tag{6.20}$$

for $p = (1/2) + \delta$. However, we can calculate the expectation in (6.20) exactly. Since $R_1 - L_1 \stackrel{\mathcal{L}}{=} 1 - U$, where U has a unif$(0,1)$ distribution,

$$\begin{aligned}\mathbb{E}G_2^p &= \sum_{i=1}^{\infty} i^p \mathbb{E}[(1-U)U^{i-1}] \\ &= \sum_{i=1}^{\infty} \frac{i^p}{i(i+1)},\end{aligned}$$

which is finite because $p < 1$. □

6.3 Convergence of moments for `QuickVal` residual

The main result of this section is that, under suitable tameness assumptions for the cost function, the moments of the normalized `QuickVal` residual converge to those of its limiting distribution.

Theorem 6.14. *Let $p \in [2, \infty)$. Suppose that the cost function β is ϵ-tame with $\epsilon < 1/p$. Then the moments of orders $\leq p$ for the normalized* `QuickVal` *residual*

$$\sqrt{n}\left(\frac{S_n^{(\mathrm{V})}}{n} - S\right)$$

converge to the corresponding moments of the limit-law random variable $\sigma_\infty Z$.

Remark 6.15. We will prove Theorem 6.14 using the second assertion in [6, Theorem 4.5.2]. Use of the first assertion in that theorem shows that, for all real $r \in [1, p]$, we also have convergence of rth absolute moments.

As mentioned in Remark 6.15, we prove Theorem 6.14 using [6, Theorem 4.5.2] by proving that, for some $q > p$, the L^q-norms of the normalized `QuickVal` residuals are bounded in n. Choosing q arbitrarily from the nonempty interval $[2, 1/\epsilon)$ and using the triangle inequality for L^q-norm, we do this by showing (in Lemmas 6.16 and 6.17, respectively) that the same L^q-boundedness holds for each of the following two sequences:

$$
\begin{aligned}
W_n &= \frac{1}{\sqrt{n}}\left[S_n^{(\mathrm{V})} - \sum_{k=1}^{\infty}(n-\tau_k)^+ I_k\right] = \frac{1}{\sqrt{n}}\sum_{k=1}^{\infty}\left[S_{k,n} - (n-\tau_k)^+ I_k\right] \\
&= \frac{1}{\sqrt{n}}\sum_{k=1}^{\infty}\sum_{\tau_k < i \le n}(X_{k,i} - I_k),
\end{aligned}
$$

and the sequence previously treated in Proposition 6.13:

$$
\widehat{W}_n := \frac{1}{\sqrt{n}}\sum_{k=1}^{\infty}\left[n - (n-\tau_k)^+\right] I_k.
$$

Lemma 6.16. *Let $q \in [2, \infty)$, and suppose that the cost function β is ϵ-tame with $0 \le \epsilon < 1/q$. Then the sequence (W_n) is L^q-bounded.*

Proof. This is straightforward. We proceed as at (6.10), except that we use triangle inequality for L^q-norm rather than for L^2-norm:

$$
\|W_n\|_q \le \frac{1}{\sqrt{n}}\sum_{k=1}^{\infty}\left\|\sum_{\tau_k < i \le n}(X_{k,i} - I_k)\right\|_q .
$$

To bound the L^q-norm on the right, we employ Rosenthal's inequality [33] conditionally given C_k to find

$$\left\| \sum_{\tau_k < i \le n} (X_{k,i} - I_k) \right\|_q^q \le c_q \left[(n-\tau_k)^+ \|X_k - I_k\|_q^q + \left((n-\tau_k)^+\right)^{q/2} \|X_k - I_k\|_2^2 \right]$$
$$\le c_q \left[n\|X_k - I_k\|_q^q + n^{q/2}\|X_k - I_k\|_2^2 \right],$$

and so, by Lemma 6.12,

$$\left\| \sum_{\tau_k < i \le n} (X_{k,i} - I_k) \right\|_q \le c_q^{1/q} \left[n^{1/q}\|X_k - I_k\|_q + n^{1/2}\|X_k - I_k\|_2^{2/q} \right].$$

But by the argument at (6.11) we have

$$\|X_k - I_k\|_2^2 \le \mathbb{E} I_{2,k},$$

and

$$\|X_k - I_k\|_q \le \|X_k\|_q + \|I_k\|_q = (\mathbb{E} I_{q,k})^{1/q} + \|I_k\|_q.$$

by again conditioning on C_k to obtain the equality here. Consider a generalization of the definition of $I_{2,k,k}$ given in (6.3):

$$I_{q,k} := \mathbb{E}\left[X_k^q \,|\, C_k \right] = \int_{L_{k-1}}^{R_{k-1}} \beta^q(u, U_{\tau_k})\, du.$$

Therefore

$$\left\| \sum_{\tau_k < i \le n} (X_{k,i} - I_k) \right\|_q \le c_q^{1/q} \left\{ n^{1/q} \left[(\mathbb{E} I_{q,k})^{1/q} + \|I_k\|_q \right] + n^{1/2} (\mathbb{E} I_{2,k})^{1/q} \right\}.$$

Three applications of Lemma 2.4 do the rest. □

Lemma 6.17. *Suppose that the cost function β is ϵ-tame with $0 \le \epsilon < 1/2$. Then the sequence $(\widehat{W}_n)$ is L^q-bounded for every $q < \infty$.*

Proof. We may and do suppose $q \ge 2$. We begin as in the proof of Proposition 6.13, except that there is now no harm in choosing $\delta = 0$. So it is sufficient to prove that

$$\sum_{k=1}^{\infty} \left\| \tau_k^{1/2} I_k \right\|_q < \infty.$$

We follow the proof of Proposition 6.13 to a large extent; in particular, what we will show is that all the terms in this sum are finite and that, for sufficiently large K, the series $\sum_{k=K}^{\infty}$ converges. As in the proof of Proposition 6.13, we utilize the bound

$$I_k \le \frac{2^{\epsilon} c}{1-\epsilon} (R_{k-1} - L_{k-1})^{1-\epsilon},$$

which requires only ϵ-tameness with $\epsilon < 1$. Then we proceed much the same way as at (6.17), but now substituting convexity of qth power for use of Lemma 6.12:

$$\begin{aligned} \mathbb{E}\left[\left(\tau_k^{1/2} I_k \right)^q \middle| \mathcal{F}_k \right] &\le C_1^q \, \mathbb{E}\left[\left(\sum_{i=0}^{k-1} G_{k,i} \right)^{q/2} (R_{k-1} - L_{k-1})^{q(1-\epsilon)} \middle| L_{k-1}, R_{k-1} \right] \\ &\le C_1^q \, (R_{k-1} - L_{k-1})^{q(1-\epsilon)} k^{(q/2)-1} \, \mathbb{E}\left[\sum_{i=0}^{k-1} G_{k,i}^{q/2} \middle| L_{k-1}, R_{k-1} \right] \\ &\le C_1^q k^{q/2} (R_{k-1} - L_{k-1})^{q(1-\epsilon)} \, \mathbb{E}\left[G_k^{q/2} \middle| L_{k-1}, R_{k-1} \right], \end{aligned} \tag{6.21}$$

where, as before, $C_1 = 2^{\epsilon} c/(1-\epsilon)$.

Arguing from here just as in the proof of Proposition 6.13, we find

$$\mathbb{E}\left[G_k^{q/2} \middle| L_{k-1}, R_{k-1} \right] \le C_2 \sum_{j=0}^{\infty} (R_{k-1} - L_{k-1})^{j-(q/2)}$$

where $C_2 := \max(\Gamma(1 + (q/2)), C_{q/2,\eta})$. (See the proof of Proposition 6.13 for the definition of $C_{q/2,\eta}$.) Therefore, with $C_3 := C_1^{q/2} C_2$, we have

$$\mathbb{E}\left[\left(\tau_k^{1/2} I_k\right)^q \middle| \mathcal{F}_k\right] \le C_3\, k^{q/2} \sum_{j=0}^{\infty} (R_{k-1} - L_{k-1})^{j+q(1-\epsilon)-(q/2)}.$$

By [12, Lemma 3.5], we have (using our assumption $\epsilon < 1/2$ for the $j = 0$ term)

$$\mathbb{E}\left(\tau_k^{1/2} I_k\right)^q \le C_3\, k^{q/2} \sum_{j=0}^{\infty} \gamma_{j,q,\epsilon}^{k-1}, \tag{6.22}$$

where

$$\gamma_{j,q,\epsilon} := \frac{2 - 2^{-[j+q(1-\epsilon)-(q/2)]}}{j + q(1-\epsilon) - (q/2) + 1} \in (0, 1)$$

decreases in j and vanishes in the limit as $j \to \infty$. Therefore, taking qth roots and using Lemma 6.12,

$$\left\|\tau_k^{1/2} I_k\right\|_q \le C_3^{q/2} k^{1/2} \sum_{j=0}^{\infty} \gamma_{j,q,\epsilon}^{(k-1)/q}.$$

If we bound the factor $k^{1/2}$ here by k and then sum the right side over $k \ge K$, the result is

$$C_3^{q/2} \sum_{j=0}^{\infty} \left[(K-1)\frac{\Gamma_j^{K-1}}{1-\Gamma_j} + \frac{\Gamma_j^{K-1}}{(1-\Gamma_j)^2}\right] \le C_3^{q/2} K \sum_{j=0}^{\infty} \frac{\Gamma_j^{K-1}}{(1-\Gamma_j)^2},$$

where

$$\Gamma_j \equiv \Gamma_{j,q,\epsilon} := \gamma_{j,q,\epsilon}^{1/q} \in (0, 1),$$

like $\gamma_{j,q,\epsilon}$, decreases in j and vanishes in the limit as $j \to \infty$. Since $\Gamma_j < (2/j)^{1/q}$, it follows if we take $K \ge 2q + 1$ that

$$\sum_{k=K}^{\infty} \left\|\tau_k^{1/2} I_k\right\|_q < \infty.$$

It remains to show that $\left\|\tau_k^{1/2} I_k\right\|_q < \infty$ for every k. For this we use (6.22) to note, since $0 < \gamma_{j,q,\epsilon} < 2/j$, that it clearly suffices to consider the cases $k = 1$ and $k = 2$. When $k = 1$ we have $\tau_1 = 1$ and hence $\left\|\tau_1^{1/2} I_1\right\|_q = \|I_1\|_q \le C_1 < \infty$. Applying (6.21) when $k = 2$ gives

$$\mathbb{E}\left[\left(\tau_2^{1/2} I_2\right)^q \middle| \mathcal{F}_2\right] \le C_1^q 2^{q/2} (R_1 - L_1)^{q(1-\epsilon)}\, \mathbb{E}\Big[G_2^{q/2} \Big| L_1, R_1\Big],$$

and we can exactly compute

$$\begin{aligned}
&\mathbb{E}\left\{(R_1 - L_1)^{q(1-\epsilon)}\, \mathbb{E}\Big[G_2^{q/2} \Big| L_1, R_1\Big]\right\} \\
&= \mathbb{E}\left\{(R_1 - L_1)^{q(1-\epsilon)} \sum_{i=1}^{\infty} i^{q/2} (R_1 - L_1)[1 - (R_1 - L_1)]^{i-1}\right\} \\
&= \sum_{i=1}^{\infty} i^{q/2} \mathbb{E}\left[U^{i-1}(1-U)^{q(1-\epsilon)+1}\right] = \sum_{i=1}^{\infty} i^{q/2} B(i, q(1-\epsilon)+2)
\end{aligned}$$

where $U \sim \mathrm{unif}(0,1)$. Each of the terms in this last sum is finite, and by Stirling's formula the ith term equals $(1+o(1))i^{-[2+((1/2)-\epsilon)q]} = o(i^{-2})$ as $i \to \infty$, so the sum converges. Hence $\|\tau_2^{1/2} I_2\|_q < \infty$. $\square$

Chapter 7

Moments of `QuickMin` Residual

In this chapter we describe the approach, involving the contraction method for inspiration and the method of moments for proof, we initially took in trying to establish a limiting distribution for the `QuickVal` residual in the special case of `QuickMin` with key-comparisons cost. It turns out that, for this approach, we must consider the `QuickMin` limit and the residual from it *bivariately*.

In Section 7.1 we motivate and treat a certain bivariate distributional fixed-point equation (7.4), which we show in Section 7.2 has a unique solution (call it F), and we provide a representation of F as the distribution of a certain bivariate random infinite series. We argue that this F is a natural candidate for the limiting joint distribution of Y (the limit of the normalized cost Y_n) and the residual $Y_n - Y$, and indeed in Remark 7.2 we show that the second marginal of F coincides with the limit `QuickMin` residual distribution from Theorem 6.4, a theorem proved in Chapter 6 by quite different means. (Both of these distributions are scale mixtures of centered Gaussians, so it is enough in Remark 7.2 that we show that their scale-mixing distributions are equal.)

In Section 7.3 we derive a recurrence (in the orders of the moments) for the (mixed) moments of F. In Section 7.4 we prove that the mixed moments for $(Y, Y_n - Y)$ converge to the corresponding moments of F. It was eventually discovered by Jim Fill [personal communication] that, unfortunately, the limit residual `QuickMin` distribution is *not* uniquely determined by its moments; so the method-of-moments approach is ultimately unsuccessful, unlike for `QuickSort` [16]. We nevertheless find this approach instructive, since (a) the unsuccessful approach reinforced our belief in convergence in distribution for the `QuickVal` residual and motivated us to search for the successful approach of Chapter 6, and (b) it does yield a rather *direct* proof of convergence of moments for the residual in the special case of `QuickMin` with key-comparisons cost.

7.1 The contraction method and a fixed-point equation

Naïvely following contraction-method-like arguments we can define a candidate limit distribution for `QuickMin` residual cost as the solution X to the distributional equation (7.4) below. To explain, recall (5.5), which we repeat here:

$$\begin{aligned} Y_n - Y &= \frac{I_n+1}{n+1} Y_{n,0} - UY^{(0)} + C_n(I_n)\frac{n}{n+1} - C(U) \\ &= \frac{I_n+1}{n+1}\left(Y_{n,0} - Y^{(0)}\right) + \left(\frac{I_n+1}{n+1} - U\right) Y^{(0)} + C_n(I_n)\frac{n}{n+1} - C(U), \quad (7.1) \end{aligned}$$

where $C_n(i) = n^{-1}(n-1+\mu_i-\mu_n)$ and $C(x) = 2x-1$. Multiplying (7.1) by $\sqrt{n+1}$ gives

$$\begin{aligned} \sqrt{n+1}(Y_n - Y) = \sqrt{\frac{I_n+1}{n+1}}\sqrt{I_n+1}\left(Y_{n,0} - Y^{(0)}\right) + \sqrt{n+1}\left(\frac{I_n+1}{n+1} - U\right) Y^{(0)} \\ + \sqrt{n+1}\left[C_n(I_n)\frac{n}{n+1} - C(U)\right]. \quad (7.2) \end{aligned}$$

Conditionally given U, the number I_n of seeds that fall to the left of the root follows the binomial(n, U) distribution; hence as $n \to \infty$, the central limit theorem applied conditionally implies

$$\sqrt{n+1}\left(\frac{I_n+1}{n+1} - U\right) \xrightarrow{\mathcal{L}} \sqrt{U(1-U)}Z,$$

where Z has a standard normal distribution. If we assume that $\sqrt{n+1}(Y_n - Y) \xrightarrow{\mathcal{L}} X$ for some random variable, then naïvely taking limits in (7.2) gives

$$X \stackrel{\mathcal{L}}{=} \sqrt{U}\widetilde{X} + \sqrt{U(1-U)}Z\widetilde{Y}^{(0)} + 2\sqrt{U(1-U)}Z, \quad (7.3)$$

where, on the right, $\widetilde{X} \stackrel{\mathcal{L}}{=} X$, $U \sim \text{Unif}(0,1)$, $Z \sim \text{Normal}(0,1)$, and $\widetilde{Y}^{(0)} \stackrel{\mathcal{L}}{=} Y^{(0)} \stackrel{\mathcal{L}}{=} Y$ is a centered Dickman random variable, and U, Z, and $(\widetilde{X}, \widetilde{Y}^{(0)})$ are independent.

Because we anticipate that X and Y are dependent, with the same joint distribution as $\widetilde{X}$ and $\widetilde{Y}^{(0)}$, we must study X and Y *bivariately.* The same naïve argument as for (7.3) yields the following conjecture, with the notation as described above and $D := Y + 2$ and $\widetilde{D} := \widetilde{Y}^{(0)} + 2$ having *uncentered* Dickman distributions: $(\sqrt{n+1}(Y_n - Y), D) \xrightarrow{\mathcal{L}} (X, D)$, where

$$\begin{pmatrix} X \\ D \end{pmatrix} \stackrel{\mathcal{L}}{=} \begin{pmatrix} \sqrt{U} & \sqrt{U(1-U)}Z \\ 0 & U \end{pmatrix} \begin{pmatrix} \widetilde{X} \\ \widetilde{D} \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{7.4}$$

i.e.,

$$\begin{pmatrix} X \\ D \end{pmatrix} \stackrel{\mathcal{L}}{=} M \begin{pmatrix} \widetilde{X} \\ \widetilde{D} \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Here we have defined

$$M := \begin{pmatrix} \sqrt{U} & \sqrt{U(1-U)}Z \\ 0 & U \end{pmatrix}, \tag{7.5}$$

and we note that M and $(\widetilde{X}, \widetilde{D})$ are independent.

We will show in Section 7.2 that there exists a unique distribution that solves (7.4).

7.2 Finiteness of the series representation of the limit distribution

.

Repeated application of (7.4) gives the following representation of the joint distribution of the random variables X and D, where, independently of $(\widetilde{X}, \widetilde{D})$ having the same joint distribution as (X, D), the matrices $M_1, M_2, \ldots, M_n$ are independent copies of the random matrix M defined previously and $e_2 = (0,1)^T$, the second-coordinate vector:

$$\begin{pmatrix} X \\ D \end{pmatrix} \stackrel{\mathcal{L}}{=} M_1 \cdots M_n \begin{pmatrix} \widetilde{X} \\ \widetilde{D} \end{pmatrix} + \left(\sum_{j=0}^{n-1} M_1 \cdots M_j \right) e_2,$$

that is,

$$\begin{pmatrix} X \\ D \end{pmatrix} \stackrel{\mathcal{L}}{=} R_n \begin{pmatrix} \widetilde{X} \\ \widetilde{D} \end{pmatrix} + \left(\sum_{j=0}^{n-1} R_j \right) e_2, \tag{7.6}$$

with $R_n := M_1 \cdots M_n$. Here, writing

$$M_k = \begin{pmatrix} \sqrt{U_k} & \sqrt{U_k(1-U_k)} Z_k \\ 0 & U_k \end{pmatrix}$$

as at (7.5), the matrix R_n has the form

$$R_n = \begin{pmatrix} \prod_{m=1}^n \sqrt{U_m} & R_n(1,2) \\ 0 & \prod_{m=1}^n U_m \end{pmatrix}.$$

The entry of R_n in the first row and second column is given by

$$R_n(1,2) = \sum_{k=1}^n \left(\prod_{m=1}^k \sqrt{U_m} \right) \sqrt{1-U_k} Z_k \left(\prod_{m=k+1}^n U_m \right),$$

Notice that since $Z_1, Z_2, \ldots, Z_n$ are independent standard normals, conditionally given $\mathcal{F}_n$ (defined to be the σ-algebra generated by $U_1, U_2, \ldots, U_n$) we have

$$R_n(1,2) \sim \text{Normal} \left(0, \sum_{k=1}^n \left[(1-U_k) \left(\prod_{m=1}^k U_m \right) \left(\prod_{m=k+1}^n U_m^2 \right) \right] \right).$$

We now look at how (7.6) behaves as $n \to \infty$ by considering R_n. Since the diagonal entries of R_n clearly converge to zero almost surely, it suffices to consider

$R_n(1,2)$. We can bound the conditional variance of $R_n(1,2)$ given $\mathcal{F}_n$ as follows:

$$\begin{aligned}\operatorname{Var}(R_n(1,2)|\mathcal{F}_n) &\le \sum_{k=1}^{n}\left[(1-U_k)\left(\prod_{m=1}^{k}U_m\right)\left(\prod_{m=k+1}^{n}U_m\right)\right]\\ &= \left(\prod_{m=1}^{n}U_m\right)\sum_{k=1}^{n}(1-U_k)\\ &\le n\prod_{m=1}^{n}U_m.\end{aligned}$$

Therefore, noting that the U_m are independent, the law of total variance gives

$$\begin{aligned}\operatorname{Var} R_n(1,2) &\le n\mathbb{E}\prod_{m=1}^{n}U_m\\ &= n\prod_{m=1}^{n}\mathbb{E}U_m\\ &= \frac{n}{2^n}.\end{aligned}$$

Since $(\|R_n(1,2)\|_2)$ is a summable sequence,

$$\sum_{n=1}^{\infty}|R_n(1,2)| < \infty \qquad \text{a.s.} \tag{7.7}$$

(Indeed, the series here has finite expectation.)

We can conclude that $|||R_n||| \to 0$ a.s., where $|||\cdot|||$ is any matrix norm. Therefore, (7.7) implies that

$$\begin{pmatrix} X \\ D \end{pmatrix} \stackrel{\mathcal{L}}{=} \left(\sum_{j=0}^{\infty}R_j\right)e_2. \tag{7.8}$$

Of particular interest is the component corresponding to X, which is

$$X \stackrel{\mathcal{L}}{=} \sum_{j=0}^{\infty}R_j(1,2) = \sum_{j=0}^{\infty}\sum_{k=1}^{j}\left(\prod_{m=1}^{k}\sqrt{U_m}\right)\sqrt{1-U_k}Z_k\left(\prod_{m=k+1}^{n}U_m\right). \tag{7.9}$$

Remark 7.1. We have derived (7.8) from (7.4). Conversely, (7.8) satisfies (7.4). Thus we have proved that (7.8) gives the unique distribution solving (7.4).

Remark 7.2. Both (7.9) and Theorem 6.4 [with the definition (6.6)] specify scale mixtures of centered Gaussians. From (7.9), the squared-scale random variable in the moments-approach is

$$A := \sum_{k=1}^{\infty} (1 - U_k) \prod_{m=1}^{k} U_m \left(\sum_{j=k}^{\infty} \prod_{m=k+1}^{j} U_m \right)^2 .$$

On the other hand, from (6.6) the squared-scale random variable in the direct CLT-approach is

$$\sum_{k=1}^{\infty} (1 - U_{\tau_k}) U_{\tau_k} + 2 \sum_{k=1}^{\infty} \sum_{\ell=k+1}^{\infty} (1 - U_{\tau_k}) U_{\tau_\ell},$$

which (as noted in Remark 6.6) has the same distribution as

$$B := \sum_{k=1}^{\infty} \left(1 - \prod_{i=1}^{k} U_i \right) \left(\prod_{j=1}^{k} U_j \right) + 2 \sum_{k=1}^{\infty} \sum_{\ell=k+1}^{\infty} \left(1 - \prod_{i=1}^{k} U_i \right) \left(\prod_{j=1}^{\ell} U_j \right).$$

We claim that $A = B$. Here is a proof, using the shorthand $\Pi_k := \prod_{m=1}^{k} U_m$. Noting that all terms in all sums below are positive, we have

$$
\begin{aligned}
A &= \sum_{k=1}^{\infty}(1-U_k)\Pi_k\left(\sum_{j=k}^{\infty}\frac{\Pi_j}{\Pi_k}\right)^2 \\
&= \sum_{k=1}^{\infty}\frac{1-U_k}{\Pi_k}\left(\sum_{j=k}^{\infty}\Pi_j\right)^2 \\
&= \sum_{k=1}^{\infty}\left(\frac{1}{\Pi_k}-\frac{1}{\Pi_{k-1}}\right)\left(\sum_{j=k}^{\infty}\Pi_j\right)^2 \\
&= \sum_{k=1}^{\infty}\left(\frac{1}{\Pi_k}-\frac{1}{\Pi_{k-1}}\right)\sum_{\ell=k}^{\infty}\Pi_\ell\sum_{m=k}^{\infty}\Pi_m \\
&= \sum_{\ell=1}^{\infty}\sum_{m=1}^{\infty}\Pi_\ell\,\Pi_m\sum_{k=1}^{\ell\wedge m}\left(\frac{1}{\Pi_k}-\frac{1}{\Pi_{k-1}}\right) \\
&= \sum_{\ell=1}^{\infty}\sum_{m=1}^{\infty}\Pi_\ell\,\Pi_m\left(\frac{1}{\Pi_{\ell\wedge m}}-\frac{1}{\Pi_0}\right) \\
&= \sum_{\ell=1}^{\infty}\sum_{m=1}^{\infty}\Pi_\ell\,\Pi_m\left(\frac{1}{\Pi_{\ell\wedge m}}-1\right) \\
&= \sum_{\ell=1}^{\infty}\Pi_\ell(1-\Pi_\ell)+2\sum_{\ell=1}^{\infty}\sum_{m=\ell+1}^{\infty}\Pi_m(1-\Pi_\ell)=B.
\end{aligned}
$$

7.3 Recurrence for moments of limit `QuickMin` residual

Denote the joint moment of order (k,ℓ) for (X,D) in (7.8)/(7.4), with X having the candidate limit distribution for the residual and D distributed Dickman, by

$$\alpha_{k,\ell} := \mathbb{E}\left[X^k D^\ell\right].$$

The remainder of this section is devoted to deriving a recurrence for $\alpha_{k,\ell}$. Let

$$\beta_{x,j} := \mathbb{E}\left[U^x\left(\sqrt{U(1-U)}Z\right)^j\right].$$

Then

$$\beta_{x,j} = \begin{cases} 0 & \text{if } j \text{ is odd} \\ [(j-1)!!]\, B(x+(j/2)+1,(j/2)+1) & \text{if } j \text{ is even} \end{cases}$$

where B is the beta function:

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\mathrm{d}x,$$

and for any odd positive integer

$$(2m-1)!! = (2m-1)(2m-3)\cdots(3)(1)$$

is the product of all odd positive integers no larger than $2m-1$. For notational simplicity, we define $(-1)!! := 1$ when $j=0$. Then a recurrence for $\alpha_{k,\ell}$ can be derived from the expansion

$$\begin{aligned}
\alpha_{k,\ell} = \mathbb{E}X^kD^\ell &= \mathbb{E}\left[\left(\sqrt{U}\widetilde{X} + \sqrt{U(1-U)}Z\widetilde{D}\right)^k\left(U\widetilde{D}+1\right)^\ell\right] \\
&= \sum_{i=0}^{\ell}\binom{\ell}{i}\mathbb{E}\left[\left(\sqrt{U}\widetilde{X} + \sqrt{U(1-U)}Z\widetilde{D}\right)^k\left(U\widetilde{D}\right)^i\right] \\
&= \sum_{i=0}^{\ell}\sum_{j=0}^{k}\binom{\ell}{i}\binom{k}{j}\mathbb{E}\left[\left(\sqrt{U}\widetilde{X}\right)^j\left(\sqrt{U(1-U)}Z\widetilde{D}\right)^{k-j}\left(U\widetilde{D}\right)^i\right] \\
&= \sum_{i=0}^{\ell}\sum_{j=0}^{k}\binom{\ell}{i}\binom{k}{j}\mathbb{E}\left[\widetilde{X}^j\widetilde{D}^{i+k-j}\right]\mathbb{E}\left[U^{i+(j/2)}\left(\sqrt{U(1-U)}Z\right)^{k-j}\right].
\end{aligned}$$

Indeed, recalling the definitions of $\alpha_{a,b}$ and $\beta_{c,d}$ gives

$$\alpha_{k,\ell} = \sum_{i=0}^{\ell}\sum_{j=0}^{k}\binom{\ell}{i}\binom{k}{j}\alpha_{j,i+k-j}\beta_{i+(j/2),k-j}.$$

Noting that

$$\beta_{\ell+(k/2),0} = \frac{1}{\ell + (k/2) + 1},$$

and defining (for notational simplicity) $I := \{(i,j) : 0 \le i \le \ell, 0 \le j \le k\} \setminus \{(\ell, k)\}$ gives

$$\alpha_{k,\ell} = \left(1 + \frac{2}{k+2\ell}\right) \sum_{(i,j)\in I} \binom{\ell}{i}\binom{k}{j} \alpha_{j,i+k-j}\beta_{i+(j/2),k-j}. \tag{7.10}$$

7.4 Recurrence for moments of QuickMin residual

Recall from Chapter 5 the definitions [see (5.1)–(5.3)] of Y_n, Y, and I_n. For k and ℓ fixed define

$$a_n^{k,\ell} := \mathbb{E}\left\{(Y+2)^\ell \left[\sqrt{n+1}\,(Y_n - Y)\right]^k\right\}, \tag{7.11}$$

$$J_n := \frac{I_n + 1}{n+1},$$

and

$$\eta_n := \frac{1}{n+1}\left(H_n - H_{I_n} - 1\right), \tag{7.12}$$

where H_n denotes the n-th harmonic number; and define

$$\gamma(n,i,\mathbf{j}) := \mathbb{E}\left\{U^i J_n^{j_1/2} \left[\sqrt{n+1}\,(J_n - U)\right]^{j_2} a_{I_n}^{j_1, j_2+i} \eta_n^{j_3}\right\},$$

where $\mathbf{j} = (j_1, j_2, j_3) \ge \mathbf{0}$ and $j_1 + j_2 + j_3 = k$. (Note that in the definition of $\mathbf{j}$ we omit the dependence on k.) When $j_3 = 0$, we have $j_2 = k - j_1$ and so in this case, we will omit the parentheses and the j_3 argument:

$$\gamma(n,i,j,k-j) \ := \gamma(n,i,(j,k-j,0)).$$

Our ultimate goal is a proof that for all k and ℓ the mixed moments $a_n^{k,\ell}$ converge to the moments $\alpha_{k,\ell}$ defined in Section 7.3. We accomplish this in two steps: (1) We derive a recurrence for $a_n^{k,\ell}$ in Proposition 7.9, and (2) we show under a suitable tameness condition that this recurrence converges (in a suitable sense) to (7.10).

Observe that, conditionally given I_n, the random variable U has a Beta$(I_n+1, n-I_n)$ distribution. Our technique is to approximate U by its conditional expectation given I_n, which is J_n, and to approximate the difference between U and J_n by a suitably scaled normal random variable. However, first we express $a_n^{k,\ell}$ in terms of $\gamma(n,i,\mathbf{j})$:

Lemma 7.3. *For any choice of* $(k,\ell,n) \in \mathbb{Z}^3_{\geq 0}$, *we have*

$$a_n^{k,\ell} = \sum_{i=0}^{\ell} \sum_{\boldsymbol{j}: j_1+j_2+j_3=k} \binom{\ell}{i}\binom{k}{\boldsymbol{j}} \gamma(n,i,\boldsymbol{j}),$$

Proof. Recall from (5.2)–(5.4) that $\widetilde{D} = Y^{(0)} + 2$ and $Y + 2 \stackrel{\mathcal{L}}{=} U\widetilde{D} + 1$. We now recall an equation for $Y_n - Y$ (which we scale by $\sqrt{n+1}$) derived in Section 5 combining (5.6) with (5.8):

$$\begin{aligned}\sqrt{n+1}\,(Y_n - Y) &= \left(\frac{I_n+1}{n+1}\right) Y_{n,0} - UY^{(0)} + C_n(I_n)\frac{n}{n+1} - C(U)\\ &= J_n^{1/2}\sqrt{I_n+1}\left(Y_{n,0} - Y^{(0)}\right) + \sqrt{n+1}\,(J_n - U)\,\widetilde{D}\\ &\quad + \frac{2}{\sqrt{n+1}}\left(H_n - H_{I_n} - 1\right).\end{aligned}$$

We use our decompositions of $Y+2$ and $\sqrt{n+1}\,(Y_n - Y)$ to express these factors in terms of gamma:

$$\begin{aligned}
a_n^{k,\ell} &= \mathbb{E}\left\{\left[\sqrt{n+1}\,(Y_n - Y)\right]^k \left[U\widetilde{D}+1\right]^\ell\right\} \\
&= \sum_{i=0}^{\ell}\binom{\ell}{i}\mathbb{E}\left\{U^i\widetilde{D}^i\left[\sqrt{n+1}\,(Y_n - Y)\right]^k\right\} \\
&= \sum_{i=0}^{\ell}\binom{\ell}{i}\mathbb{E}\Bigg\{U^i\widetilde{D}^i \sum_{\mathbf{j}:j_1+j_2+j_3=k}\binom{k}{\mathbf{j}} J_n^{j_1/2}\left[\sqrt{I_n+1}\left(Y_{n,0}-Y^{(0)}\right)\right]^{j_1} \\
&\qquad \cdot\left[\sqrt{n+1}\,(J_n - U)\,\widetilde{D}\right]^{j_2}\eta_n^{j_3}\Bigg\}.
\end{aligned}$$

Recall Remark 5.3 concerning the existence of a probabilistic copy (Y_n^*) of the $(Y_n)_{n=1}^\infty$ process independent of U and I_n such that $Y_{n,0} \stackrel{\mathcal{L}}{=} Y_{I_n}^*$. Therefore we can rewrite the previous expression as

$$\begin{aligned}
a_n^{k,\ell} = &\sum_{i=0}^{\ell}\sum_{\mathbf{j}:j_1+j_2+j_3=k}\binom{\ell}{i}\binom{k}{\mathbf{j}}\mathbb{E}\Bigg\{U^i\widetilde{D}^i J_n^{j_1/2}\left[\sqrt{I_n+1}\left(Y_{I_n}^*-Y^*\right)\right]^{j_1} \\
&\cdot\left[\sqrt{n+1}\,(J_n - U)\,(Y^*+2)\right]^{j_2}\eta_n^{j_3}\Bigg\}
\end{aligned}\tag{7.13}$$

Define

$$A_n := \sqrt{I_n+1}\left(Y_{I_n}^* - Y^*\right).$$

To finish the proof of the lemma, we condition on U and I_n and need to show

$$\begin{aligned}
\gamma(n,i,\mathbf{j}) &= \mathbb{E}\left\{U^i J_n^{j_1/2}\left[\sqrt{n+1}\,(J_n - U)\right]^{j_2} a_{I_n}^{j_1,j_2+i}\eta_n^{j_3}\right\} \\
&= \mathbb{E}\left\{U^i J_n^{j_1/2}\left[\sqrt{n+1}\,(J_n - U)\right]^{j_2}\eta_n^{j_3}\mathbb{E}\left[A_n^{j_1}(Y^*+2)^{j_2+i}\,|U, I_n\right]\right\}.
\end{aligned}$$

However, this follows from the fact that (A_n, Y^*) is conditionally independent of U given I_n. $\square$

Define

$$b_n^{k,\ell} := \mathbb{E}\left|(Y+2)^\ell \left[\sqrt{n+1}\,(Y_n - Y)\right]^k\right|.$$

Note that $b_n^{k,\ell}$ is defined precisely like $a_n^{k,\ell}$ (7.11), except for the absolute value inside the expectation. Therefore, we immediately have $|a_n^{k,\ell}| \le b_n^{k,\ell}$. The next lemma follows mutatis mutandis as in Lemma 7.3.

Lemma 7.4. *For any choice of* $(k, \ell, n) \in \mathbb{Z}^3_{\ge 0}$*, we have*

$$b_n^{k,\ell} \le \sum_{i=0}^{\ell} \sum_{\boldsymbol{j}: j_1+j_2+j_3=k} \binom{\ell}{i}\binom{k}{\boldsymbol{j}} \mathbb{E}\left\{U^i J_n^{j_1/2} W_n^{j_2} b_{I_n}^{j_1, j_2+i} \sigma_n^{j_3}\right\},$$

where $\sigma_n := |\eta_n|$*, and* $W_n := \sqrt{n+1}\,|J_n - U|$*.*

Lemma 7.5. *Let* $B \sim \mathrm{Beta}(m+1, n-m)$*. Then for each* $r \ge 2$ *we have*

$$\mathbb{E}\left|B - \frac{m+1}{n+1}\right|^r = O(n^{-r/2})$$

uniformly in $m \in \{0, 1, \ldots, n-1\}$*.*

Proof. Let $T_1, \ldots, T_n$ be independent exponential random variables with unit rate. If we define $G_k := \sum_{i=1}^k T_i$, then

$$B \overset{\mathcal{L}}{=} \frac{G_m}{G_n}.$$

We will show that

$$\left\|\frac{G_m}{G_n} - \frac{m+1}{n+1}\right\|_r = O(n^{-1/2}),$$

where $\|\cdot\|_r := [\mathbb{E}|\cdot|^r]^{1/r}$ is the L^r-norm. We can bound this norm using Minkowski's inequality:

$$\begin{aligned}\left\|\frac{G_m}{G_n} - \frac{m+1}{n+1}\right\|_r &\le \left\|\frac{1}{n+1}(G_m - (m+1))\right\|_r + \left\|G_m\left(\frac{1}{G_n} - \frac{1}{n+1}\right)\right\|_r \\ &\le \left\|\frac{1}{n+1}(G_m - (m+1))\right\|_r + \left\|G_n\left(\frac{1}{G_n} - \frac{1}{n+1}\right)\right\|_r \\ &= \frac{1}{n+1}\left[\|G_m - (m+1)\|_r + \|G_n - (n+1)\|_r\right]. \qquad (7.14)\end{aligned}$$

It suffices to show that each norm in (7.14) is $O(n^{1/2})$. We can bound each term with Rosenthal's inequality [33], which gives

$$\begin{aligned}\|G_k - (k+1)\|_r &\le c_r \max\left\{\left(\sum_{j=1}^{k}\|X_j\|_r^r\right)^{1/r}, \left(\sum_{j=1}^{k}\|X_j\|_2^2\right)^{1/2}\right\} \\ &= c_r \max\{k^{1/r}\|X_1\|_r, k^{1/2}\|X_1\|_2\} \\ &\le c_r\|X_1\|_r\sqrt{k},\end{aligned}$$

where $X_j := T_j - 1$. □

We next prove a finite bound on $b_n^{k,\ell}$.

Proposition 7.6. *For any $(k,\ell) \in \mathbb{Z}^2_{\ge 0}$, there exist finite constants $B(k,\ell)$ such that*

$$b_n^{k,\ell} \le B(k,\ell)$$

for any n.

Proof. We prove this result through nested induction: an outer induction on $k+\ell$, a middle induction on k for fixed $k+\ell$, and an inner induction on n. For the outer and middle inductions, the base case of $k = \ell = 0$ is trivial. (Choose $B(0,0) = 1$.) For any k, ℓ, when $n = 0$, we have

$$b_0^{k,\ell} = \mathbb{E}\left(|Y+2|^\ell \, |Y|^k\right),$$

and these numbers are all finite. We will choose the bounds $B(k,\ell)$ so that

$$b_0^{k,\ell} \le B(k,\ell).$$

Fix n, k, ℓ and suppose that for $m < n$, for all $s, t \in \mathbb{Z}^2_{\geq}$, there exists constants $B(s,t)$ such that

$$b_m^{s,t} \le B(s,t).$$

Suppose that for all $s', t' \ge 0$ with $s' + t' < k + \ell$, we have

$$b_n^{s',t'} \le B(s',t').$$

Suppose that for $s'' < k$, we have

$$b_n^{s'',k-s''+\ell} \le B(s'', k - s'' + \ell).$$

Applying Lemma 7.4 and grouping terms gives

$$\begin{aligned} b_n^{k,\ell} \le\ & \mathbb{E}\left[U^\ell J_n^{k/2} b_{I_n}^{k,\ell}\right] \\ & + \sum_{j=0}^{k-1} \binom{k}{j} \mathbb{E}\left[U^\ell J_n^{j/2} W_n^{k-j} b_{I_n}^{j,\ell+k-j}\right] && (7.15) \\ & + \sum_{i=0}^{\ell-1} \sum_{j=0}^{k} \binom{\ell}{i} \binom{k}{j} \mathbb{E}\left[U^i J_n^{j/2} W_n^{k-j} b_{I_n}^{j,i+k-j}\right] && (7.16) \\ & + \sum_{i=0}^{\ell} \sum_{\mathbf{j}: j_3 > 0} \binom{\ell}{i} \binom{k}{\mathbf{j}} \mathbb{E}\left[U^i J_n^{j_1/2} \sigma_n^{j_3} W_n^{j_2} b_{I_n}^{j_1,i+j_2}\right]. && (7.17) \end{aligned}$$

Since, conditionally given I_n, the random variable U has a Beta$(I_n+1, n-I_n)$ distribution, as a consequence of Lemma 7.5 there exists a bound $M(j_2)$ such that

$$\mathbb{E}\left[W_n^{j_2} \middle| I_n\right] \le M(j_2).$$

Applying the induction hypothesis on k for fixed $k+\ell$ to (7.15), and the induction hypothesis on $k+\ell$ to (7.16) and (7.17) give

$$\begin{aligned} b_n^{k,\ell} \le{}& \mathbb{E}\left[U^\ell b_{I_n}^{k,\ell}\right] \\ &+ \sum_{j=0}^{k-1} \binom{k}{j} M(k-j) B(j, \ell+k-j) \\ &+ \sum_{i=0}^{\ell-1} \sum_{j=0}^{k} \binom{\ell}{i} \binom{k}{j} M(k-j) B(j, i+k-j) \\ &+ \sum_{i=0}^{\ell} \sum_{\mathbf{j}:j_3>0} \binom{\ell}{i} \binom{k}{\mathbf{j}} M(j_2) B(j_1, i+j_2). \end{aligned} \tag{7.18}$$

If we define

$$\begin{aligned} B'(k,\ell) :={}& \sum_{j=0}^{k-1} \binom{k}{j} M(k-j) B(j, \ell+k-j) \\ &+ \sum_{i=0}^{\ell-1} \sum_{j=0}^{k} \binom{\ell}{i} \binom{k}{j} M(k-j) B(j, i+k-j) \\ &+ \sum_{i=0}^{\ell} \sum_{\mathbf{j}:j_3>0} \binom{\ell}{i} \binom{k}{\mathbf{j}} M(j_2) B(j_1, i+j_2) \end{aligned}$$

and

$$B(k,\ell) := \max\left\{ b_0^{k,\ell}, \left(\frac{\ell+1}{\ell}\right) B'(k,\ell) \right\},$$

applying the induction hypothesis on n gives

$$\begin{aligned} b_n^{k,\ell} &\le (\mathbb{E}\, U^\ell) B(k,\ell) + B'(k,\ell) \\ &= \left(\frac{1}{\ell+1}\right) B(k,\ell) + B'(k,\ell) \\ &\le B(k,\ell). \end{aligned}$$

□

Define

$$\eta'(n,\ell,k) := \sum_{i=0}^{\ell} \sum_{\mathbf{j}:j_3>0} \binom{\ell}{i} \binom{k}{\mathbf{j}} \mathbb{E}\left| U^i J_n^{j_1/2} \eta_n^{j_3} V_n^{j_2} a_{I_n}^{j_1,i+j_2} \right|. \tag{7.19}$$

Corollary 7.7. *For fixed k and ℓ we have*

$$\eta'(n,\ell,k) = O\left(\frac{\log n}{n}\right).$$

Proof. From the definition (7.12) of η_n and the triangle inequality we can bound the random variable $|\eta_n|$ by $(2H_n - 1)/(n+1) = O((\log n)/n)$. Therefore, applying the bounds from Lemma 7.5 and Proposition 7.6 gives the result. □

By applying the binomial theorem to $U^i = (U - J_n + J_n)^i$, we can approximate U by J_n in $\gamma(n,i,j,k-j)$. Recall

$$\gamma(n,i,j,k-j) = \mathbb{E}\left\{ U^i J_n^{j/2} V_n^{k-j} a_{I_n}^{j,k-j+i} \right\},$$

and expand

$$\begin{aligned} U^i &= (U - J_n + J_n)^i \\ &= J_n^i + (U - J_n) \sum_{t=0}^{i-1} \binom{i}{t} (U - J_n)^{i-t-1} J_n^t. \end{aligned} \tag{7.20}$$

If we define

$$\xi(n,i,j,k-j) := \mathbb{E}\Big\{J_n^{j/2} a_{I_n}^{j,k-j+i} \mathbb{E}\Big[(n+1)^{(k-j)/2}(J_n-U)^{k-j+1} \times \sum_{t=0}^{i-1}\binom{i}{t}(U-J_n)^{i-t-1}J_n^t \Big| I_n\Big]\Big\},$$

substituting (7.20) into the definition of $\gamma(n,i,j,k-j)$ gives

$$\gamma(n,i,j,k-j) = \mathbb{E}\Big\{J_n^{j/2} a_{I_n}^{j,k-j+i} \mathbb{E}\Big[J_n^i (n+1)^{(k-j)/2}(J_n-U)^{k-j} \Big| I_n\Big]\Big\} - \xi(n,i,j,k-j). \tag{7.21}$$

Lemma 7.8. *For $(k,\ell) \in \mathbb{Z}_{\geq}^2$ and any $i \leq \ell$, and $j \leq k$, we have*

$$\xi(n,i,j,k-j) = O(n^{-1/2}).$$

Proof.

$$\begin{aligned}\xi(n,i,j,k-j) &= \mathbb{E}\Big\{J_n^{j/2} a_{I_n}^{j,k-j+i} \mathbb{E}\Big[(n+1)^{(k-j)/2}(J_n-U)^{k-j+1} \times \sum_{t=0}^{i-1}\binom{i}{t}(U-J_n)^{i-t-1}J_n^t \Big| I_n\Big]\Big\} \\ &= \frac{1}{\sqrt{n+1}}\mathbb{E}\left\{J_n^{j/2} a_{I_n}^{j,k-j+i} \mathbb{E}\left[V_n^{k-j+1}\sum_{t=0}^{i-1}\binom{i}{t}(U-J_n)^{i-t-1}J_n^t \middle| I_n\right]\right\}.\end{aligned}$$

Using $0 < J_n < 1$ and Proposition 7.6, the absolute value of

$$\sqrt{n+1}\,\xi(n,i,j,k-j)$$

is bounded above by

$$(n+1)^{(k-j+1)/2} B(j,k-j+i) \sum_{t=0}^{i-1}\binom{i}{t}\mathbb{E}\left\{\mathbb{E}\left[|J_n-U|^{(k-j+1)+(i-t-1)} \middle| I_n\right]\right\}.$$

Now using Lemma 7.5, the sum here is bounded above by a constant (depending on k and ℓ) times

$$\sum_{t=0}^{i-1} \binom{i}{t} (n+1)^{-[(k-j+1)+(i-t-1)]/2}.$$

Since $\sum_{i=0}^{\infty}(n+1)^{-i/2} = [1-(n+1)^{-1/2}]^{-1}$ decreases in n and equals $(1-2^{-1/2})^{-1} < \infty$ when $n = 1$, the result follows.

□

Recall the definition of the set I from Section 7.3:

$$I := \{(i, j) : 0 \leq i \leq \ell, 0 \leq j \leq k\} \setminus \{(\ell, k)\}.$$

We can apply Lemma 2.15 to solve the recurrence for $a_n^{k,\ell}$ in terms of "smaller" k and ℓ.

Proposition 7.9. *For* $(k, \ell) \in \mathbb{Z}_{\geq}^2$, *we have*

$$\begin{aligned} a_n^{k,\ell} = O(n^{-1/2}) &+ \sum_{(i,j)\in I} \binom{\ell}{i}\binom{k}{j} \gamma(n, i, j, k - j) \\ &+ \sum_{(i,j)\in I} \binom{\ell}{i}\binom{k}{j} \frac{1}{n} \sum_{p=0}^{n-1} \left(\frac{p+1}{n+1}\right)^{\ell+(k/2)-1} \gamma(p, i, j, k - j). \end{aligned}$$

Proof. Define

$$\begin{aligned} A_n^{k,\ell} &:= (n+1)^{\ell+(k/2)} a_n^{k,\ell} \\ B_n^{k,\ell} &:= (n+1)^{\ell+(k/2)} \left[\sum_{(i,j)\in I} \binom{\ell}{i}\binom{k}{j} \gamma(n, i, j, k - j) + \eta'(n, k, \ell) - \xi(n, \ell, k, 0) \right]. \end{aligned}$$

Then by Lemma 7.3, we can write A_n as

$$A_n^{k,\ell} = (n+1)^{\ell+(k/2)}\gamma(n,\ell,k,0) + (n+1)^{\ell+(k/2)}\left[\sum_{(i,j)\in I}\binom{\ell}{i}\binom{k}{j}\gamma(n,i,j,k-j)+\eta'(n,\ell,k)\right]. \tag{7.22}$$

Rewriting $\gamma(n,\ell,k,0)$ and applying Lemma 7.8 gives

$$\begin{aligned}\gamma(n,\ell,k,0) &= \mathbb{E}\left[U^\ell J_n^{k/2} a_{I_n}^{k,\ell}\right] \\ &= \mathbb{E}\left[J_n^{\ell+(k/2)} a_{I_n}^{k,\ell}\right] - \xi(n,\ell,k,0) \\ &= \frac{1}{n}\sum_{m=0}^{n-1}\left(\frac{m+1}{n+1}\right)^{\ell+(k/2)} a_m^{k,\ell} - \xi(n,\ell,k,0).\end{aligned} \tag{7.23}$$

Substituting (7.23) into (7.22) gives

$$A_n^{k,\ell} = \frac{1}{n}\sum_{m=0}^{n-1} A_m^{k,\ell} + B_n^{k,\ell}.$$

Lemma 2.15 immediately implies

$$A_n^{k,\ell} = A_0^{k,\ell} + B_n^{k,\ell} + \sum_{p=1}^{n-1}\frac{1}{p+1}B_p^{k,\ell}. \tag{7.24}$$

Define

$$\widetilde{B}_n^{k,\ell} := (n+1)^{\ell+(k/2)}\left[\sum_{(i,j)\in I}\binom{\ell}{i}\binom{k}{j}\gamma(n,i,j,k-j)\right].$$

It suffices to show that

$$D(n,\ell,k) := \frac{1}{(n+1)^{\ell+(k/2)}}\left[A_n^{k,\ell} - \left(\widetilde{B}_n^{k,\ell} + \sum_{p=0}^{n-1}\frac{1}{p+1}\widetilde{B}_p^{k,\ell}\right)\right] = O(n^{-1/2}).$$

Since

$$D(n,\ell,k) \le |\eta'(n,\ell,k)| + |\xi(n,\ell,k,0)| + \sum_{p=0}^{n-1} \left(\frac{1}{p+1}\right) \frac{(p+1)^{\ell+(k/2)}}{(n+1)^{\ell+(k/2)}} |\eta'(p,\ell,k) + \xi(p,\ell,k,0)|,$$

we need to show that

$$\sum_{p=0}^{n-1} \left(\frac{1}{p+1}\right) \frac{(p+1)^{\ell+(k/2)}}{(n+1)^{\ell+(k/2)}} |\eta'(p,\ell,k) + \xi(p,\ell,k,0)| = O(n^{-1/2}).$$

Indeed, using Corollary 7.7 and Lemma 7.8 we can bound the sum here by a constant (depending on k and ℓ) times

$$\sum_{p=0}^{n-1} \frac{(p+1)^{\ell+(k/2)-(3/2)}}{(n+1)^{\ell+(k/2)}} = O(n^{-1/2}).$$

□

Lemma 7.10. *For $n = 1, 2, 3, \ldots$, suppose $X_n \sim \mathrm{Beta}(\alpha_n, \beta_n)$ and define*

$$\mu_n := \frac{\alpha_n}{\alpha_n + \beta_n}, \qquad \sigma_n := \sqrt{\frac{\alpha_n \beta_n}{(\alpha_n + \beta_n)^2(\alpha_n + \beta_n + 1)}}$$

as the expected value and standard deviation of X_n. If $\alpha_n \to \infty$ and $\beta_n \to \infty$, then

$$Y_n := \sigma_n^{-1}(X_n - \mu_n) \xrightarrow{\mathcal{L}} \mathrm{N}(0,1).$$

Proof. The proof is a straightforward application of Scheffé's Theorem [4, Problem 25.9]. Denote the density of X_n by

$$f_n(x) = \frac{\Gamma(\alpha_n + \beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)} x^{\alpha_n - 1}(1-x)^{\beta_n - 1}.$$

Since $\alpha_n \to \infty$ and $\beta_n \to \infty$, Stirling's formula gives (uniformly in x)

$$\begin{aligned}
f_n(x) &= (1+o(1))\frac{\sqrt{2\pi(\alpha_n+\beta_n)}((\alpha_n+\beta_n)/e)^{\alpha_n+\beta_n}}{\sqrt{2\pi\alpha_n}(\alpha_n/e)^{\alpha_n}\sqrt{2\pi\beta_n}(\beta_n/e)^{\beta_n}}\left(\frac{\alpha_n\beta_n}{\alpha_n+\beta_n}\right)x^{\alpha_n-1}(1-x)^{\beta_n-1}\\
&= (1+o(1))\sqrt{\frac{(\alpha_n+\beta_n)^3}{2\pi\alpha_n\beta_n}}\left(x\frac{\alpha_n+\beta_n}{\alpha_n}\right)^{\alpha_n-1}\left[(1-x)\frac{\alpha_n+\beta_n}{\beta_n}\right]^{\beta_n-1}\\
&= (1+o(1))\frac{1}{\sigma_n\sqrt{2\pi}}\left(x\mu_n^{-1}\right)^{\alpha_n-1}\left[(1-x)(1-\mu_n)^{-1}\right]^{\beta_n-1}.
\end{aligned}$$

Let g_n be the density of Y_n; then (again, uniformly in y)

$$\begin{aligned}
g_n(y) &= \sigma_n f_n(\sigma_n y+\mu_n)\\
&= (1+o(1))\frac{1}{\sqrt{2\pi}}\left(1+y\frac{\sigma_n}{\mu_n}\right)^{\alpha_n-1}\left(1-y\frac{\sigma_n}{1-\mu_n}\right)^{\beta_n-1}.
\end{aligned}$$

Now fix $y \in \mathbb{R}$. We prove that $g_n(y) \to \phi(y)$, where $\phi(\cdot)$ is the standard normal density, by taking logarithms and using a Taylor's expansion. It's easy to check that $\sigma_n/\mu_n \to 0$ and $\sigma_n/(1-\mu_n) \to 0$. Therefore, $\left(1+y\frac{\sigma_n}{\mu_n}\right) \to 1$ and $\left(1-y\frac{\sigma_n}{1-\mu_n}\right) \to 1$ as $n \to \infty$, so we can first replace $\alpha_n - 1$ and $\beta_n - 1$ in the exponents by α_n and β_n, respectively. We now find

$$\begin{aligned}
\log\left[\sqrt{2\pi}g_n(y)\right] &= \alpha_n \log\left(1+y\frac{\sigma_n}{\mu_n}\right)+\beta_n\log\left(1-y\frac{\sigma_n}{1-\mu_n}\right)+o(1)\\
&= \alpha_n\left(y\frac{\sigma_n}{\mu_n}-\frac{y^2}{2}\frac{\sigma_n^2}{\mu_n^2}+O(\alpha_n^{-3/2})\right)\\
&\quad -\beta_n\left(y\frac{\sigma_n}{1-\mu_n}+\frac{y^2}{2}\left(\frac{\sigma_n}{1-\mu_n}\right)^2+O(\beta_n^{-3/2})\right)+o(1)\\
&= y\left(\frac{\alpha_n\sigma_n}{\mu_n}-\frac{\beta_n\sigma_n}{1-\mu_n}\right)-\frac{y^2}{2}\left(\frac{\alpha_n\sigma_n^2}{\mu_n^2}+\frac{\beta_n\sigma_n^2}{(1-\mu_n)^2}\right)\\
&\quad +O(\alpha_n^{-1/2})+O(\beta_n^{-1/2})+o(1)\\
&= yL_n-\frac{y^2}{2}Q_n+o(1),
\end{aligned}$$

where

$$L_n := \frac{\alpha_n \sigma_n}{\mu_n} - \frac{\beta_n \sigma_n}{1-\mu_n} \quad \text{and} \quad Q_n := \frac{\alpha_n \sigma_n^2}{\mu_n^2} + \frac{\beta_n \sigma_n^2}{(1-\mu_n)^2}.$$

Therefore it suffices to show that

$$L_n = 0 \quad \text{and} \quad Q_n \to 1$$

as $n \to \infty$. But this follows from the facts that

$$\frac{\alpha_n \sigma_n}{\mu_n} = \sqrt{\frac{\alpha_n \beta_n}{\alpha_n + \beta_n + 1}} = \frac{\beta_n \sigma_n}{1-\mu_n},$$

$$Q_n = \frac{\beta_n}{\alpha_n + \beta_n + 1} + \frac{\alpha_n}{\alpha_n + \beta_n + 1} \to 1$$

as $n \to \infty$. □

Now we place an additional restriction on (m_n), namely, that for some fixed ϵ with $0 < \epsilon < 1/2$, we require m_n to be an integer in the interval $[\epsilon n, (1-\epsilon)n]$.

Lemma 7.11. *Suppose for some ϵ satisfying $0 < \epsilon < 1/2$, we require $\epsilon n \leq m_n \leq (1-\epsilon)n$. Let $Z \sim \mathrm{Normal}(0,1)$. Then for any $r \in \mathbb{Z}_+$, we have*

$$\mathbb{E}\left[V_n^r \,|\, I_n = m_n\right] = \left[\frac{(m_n+1)(n-m_n)}{(n+1)(n+2)}\right]^{r/2} \mathbb{E}Z^r + o(1).$$

Proof. Apply Lemma 7.10 with $\alpha_n = m_n + 1$ and $\beta_n = n - m_n$ to conclude asymptotic normality for V_n. Then apply [6, Theorem 4.5.2(a)] (a result yielding convergence of moments from convergence in distribution, when the sequence of absolute moments for each fixed order is bounded), using Lemma 7.5 to verify the hypothesis of that theorem. (This is where the assumption of linear growth of both m_n and $n - m_n$ is used.) □

Remark Note that this lemma implies that

$$\mathbb{E}\left[V_n^r \middle| I_n = m\right] = \left[\frac{(m+1)(n-m)}{(n+1)(n+2)}\right]^{r/2} \mathbb{E}Z^r + o(1)$$

as $n \to \infty$, *uniformly* in m satisfying $\epsilon n \leq m \leq (1-\epsilon n)$.

Theorem 7.12. *For any* $(k, \ell) \in \mathbb{Z}_+^2$, *we have*

$$a_n^{k,\ell} \to \alpha_{k,\ell}. \tag{7.25}$$

Proof. Similar to the proof of Proposition 7.6, we use nested induction with outer induction on $k+\ell$ and inner induction on k for $k+\ell$ fixed. The base case $k=\ell=0$ is trivial. Suppose that (7.25) holds for $(s,t) \in I$. By Proposition 7.9,

$$\begin{aligned} a_n^{k,\ell} &= O(n^{-1/2}) + \sum_{(i,j)\in I} \binom{\ell}{i}\binom{k}{j} \gamma(n,i,j,k-j) \\ &+ \sum_{(i,j)\in I} \binom{\ell}{i}\binom{k}{j} \frac{1}{n} \sum_{p=0}^{n-1} \left(\frac{p+1}{n+1}\right)^{\ell+(k/2)-1} \gamma(p,i,j,k-j). \end{aligned} \tag{7.26}$$

By (7.21) and Lemma 7.8, we have

$$\gamma(p,i,j,k-j) = O(n^{-1/2}) + \frac{1}{p}\sum_{m=0}^{p-1} \mathbb{E}\left\{J_p^{i+(j/2)} a_m^{j,k-j+i} \mathbb{E}\left[V_p^{k-j} \middle| I_p = m\right]\right\}. \tag{7.27}$$

For any ϵ satisfying $0 < \epsilon < 1/2$, Proposition 7.6 and Lemma 7.5 imply the terms with $m > \lfloor (1-\epsilon)p \rfloor$ or $m < \lceil \epsilon p \rceil$ in (7.27) are $O(\epsilon)$ uniformly in p and hence in n:

$$\gamma(p,i,j,k-j) = O(\epsilon) + O(n^{-1/2}) + \frac{1}{p} \sum_{m=\lceil \epsilon p \rceil}^{\lfloor (1-\epsilon)p \rfloor} \left(\frac{m+1}{p+1}\right)^{i+(j/2)} a_m^{j,k-j+i} \mathbb{E}\left[V_p^{k-j} \middle| I_p = m\right].$$

Let Z have a standard normal distribution, and let

$$f_{m,p} := \frac{(m+1)(p-m)}{(p+1)(p+2)}.$$

Then Lemma 7.11 and the induction hypotheses imply

$$\begin{aligned}\gamma(p,i,j,k-j) &= O(\epsilon) + O(n^{-1/2}) + \frac{1}{p}\sum_{m=\lceil \epsilon p\rceil}^{\lfloor (1-\epsilon)p\rfloor} \left\{ \left(\frac{m+1}{p+1}\right)^{i+(j/2)} (\alpha^{j,k-j+i} + \delta_p^{(1)}) \right. \\ &\qquad \left. \times \left[\mathbb{E}Z^{k-j} f_{m,p}^{(k-j)/2} + \delta_p^{(2)}\right] \right\} \\ &= O(\epsilon) + o(1) + \frac{1}{p}\sum_{m=\lceil \epsilon p\rceil}^{\lfloor (1-\epsilon)p\rfloor} \left(\frac{m+1}{p+1}\right)^{i+(j/2)} \alpha^{j,k-j+i}\mathbb{E}Z^{k-j} f_{m,p}^{(k-j)/2},\end{aligned}$$

where $\delta_p^{(1)} \to 0$ and $\delta_p^{(2)} \to 0$ as $p \to \infty$ uniformly in m and the $o(1)$ term tends to zero in the limit as $p \to \infty$. For ease of notation, let

$$u := \frac{m+1}{n+1}.$$

Then for $p = n$ we have

$$\begin{aligned}\gamma(n,i,j,k-j) &= O(\epsilon) + o(1) + \alpha^{j,k-j+i}\mathbb{E}Z^{k-j}\frac{1}{n}\sum_{m=\lceil \epsilon n\rceil}^{\lfloor (1-\epsilon)n\rfloor} u^{i+(j/2)} f_{m,n}^{(k-j)/2} \\ &= O(\epsilon) + o(1) + \alpha^{j,k-j+i}\mathbb{E}Z^{k-j}\frac{1}{n}\sum_{m=\lceil \epsilon n\rceil}^{\lfloor (1-\epsilon)n\rfloor} u^{i+(k/2)} \left(\frac{n-m}{n+2}\right)^{(k-j)/2} \\ &= O(\epsilon) + o(1) + \alpha^{j,k-j+i}\mathbb{E}Z^{k-j}\frac{1}{n}\sum_{m=\lceil \epsilon n\rceil}^{\lfloor (1-\epsilon)n\rfloor} u^{i+(k/2)} \left(1-u+\delta_n^{(3)}\right)^{(k-j)/2} \\ &= O(\epsilon) + o(1) + \alpha^{j,k-j+i}\mathbb{E}Z^{k-j}\left[\frac{1}{n}\sum_{m=\lceil \epsilon n\rceil}^{\lfloor (1-\epsilon)n\rfloor} u^{i+(k/2)} (1-u)^{(k-j)/2}\right],\end{aligned}$$

where $\delta_n^{(3)} \to 0$ as $n \to \infty$ uniformly for $\lceil \epsilon n\rceil \le m \le \lfloor (1-\epsilon)n\rfloor$ and the $o(1)$ terms are as $n \to \infty$. Notice that the last expression in square brackets is a Riemann sum for the integral

$$\int_\epsilon^{1-\epsilon} x^{i+(k/2)}(1-x)^{(k-j)/2}\,\mathrm{d}x.$$

Now we consider the expression

$$R := \frac{1}{n}\sum_{p=0}^{n-1}\left(\frac{p+1}{n+1}\right)^{\ell+(k/2)-1}\gamma(p,i,j,k-j)$$

appearing in (7.26). Note that the same argument used previously to obtain an asympotic expansion of $\gamma(n,i,j,k-j)$ as $n\to\infty$ can be applied to the $\gamma(p,i,j,k-j)$ factor in the sum uniformly for $p>\lceil\epsilon n\rceil$, so that

$$R = O(\epsilon)+o(1)+\alpha^{j,k-j+i}\mathbb{E}Z^{k-j}\frac{1}{n}\sum_{p=\lceil\epsilon n\rceil}^{n-1}\left\{\left(\frac{p+1}{n+1}\right)^{\ell+(k/2)-1}\right.$$
$$\left.\times\frac{1}{p}\sum_{m=\lceil\epsilon p\rceil}^{\lfloor(1-\epsilon)p\rfloor}\left(\frac{m+1}{p+1}\right)^{i+(j/2)}\left(\frac{(m+1)(p-m)}{(p+1)(p+2)}\right)^{(k-j)/2}\right\},$$

where the $o(1)$ term is as $n\to\infty$. This double sum is a Riemann sum for the double integral

$$\int_\epsilon^1\int_\epsilon^{1-\epsilon} y^{\ell+(k/2)-1}x^{i+(k/2)}(1-x)^{(k-j)/2}\,\mathrm{d}x\,\mathrm{d}y$$
$$=\frac{1}{\ell+(k/2)}\left(1-\epsilon^{\ell+(k/2)}\right)\times\int_\epsilon^{1-\epsilon} x^{i+(k/2)}(1-x)^{(k-j)/2}\mathrm{d}x.$$

Therefore,

$$\limsup_{n\to\infty} a_n^{k,\ell}\le O(\epsilon)+\sum_{(i,j)\in I}\binom{\ell}{i}\binom{k}{j}\alpha^{j,k-j+i}\mathbb{E}Z^{k-j}\int_\epsilon^{1-\epsilon} x^{i+(k/2)}(1-x)^{(k-j)/2}\,\mathrm{d}x$$
$$+\sum_{(i,j)\in I}\binom{\ell}{i}\binom{k}{j}\alpha^{j,k-j+i}\mathbb{E}Z^{k-j}\frac{1}{\ell+(k/2)}\left(1-\epsilon^{\ell+(k/2)}\right)$$
$$\times\int_\epsilon^{1-\epsilon} x^{i+(k/2)}(1-x)^{(k-j)/2}\,\mathrm{d}x.$$

Passing to the limit as $\epsilon \downarrow 0$ then gives

$$\limsup_{n\to\infty} a_n^{k,\ell} \leq \sum_{(i,j)\in I} \binom{\ell}{i}\binom{k}{j} \alpha^{j,k-j+i} \beta_{i+(j/2),k-j} + \frac{1}{\ell + (k/2)} \sum_{(i,j)\in I} \binom{\ell}{i}\binom{k}{j} \alpha^{j,k-j+i} \beta_{i+(j/2),k-j}.$$

The reverse inequality holds similarly for lim inf, so we conclude

$$\lim_{n\to\infty} a_n^{k,\ell} = \sum_{(i,j)\in I} \binom{\ell}{i}\binom{k}{j} \alpha^{j,k-j+i} \beta_{i+(j/2),k-j} + \frac{1}{\ell + (k/2)} \sum_{(i,j)\in I} \binom{\ell}{i}\binom{k}{j} \alpha^{j,k-j+i} \beta_{i+(j/2),k-j}.$$

We can see that the expression on the right hand side is the same recurrence (7.10) that defines $\alpha^{k,\ell}$. This completes the induction. □

Bibliography

[1] N. I. Akhiezer. *The classical moment problem and some related questions in analysis.* Translated by N. Kemmer. Hafner Publishing Co., New York, 1965.

[2] P.J. Bickel and K.A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics.* Number v. 1 in Holden-Day series in probability and statistics. Prentice Hall, 2001.

[3] J. D. Biggins. Uniform convergence of martingales in the branching random walk. *Ann. Probab.*, 20(1):137–151, 01 1992.

[4] Patrick Billingsley. *Probability and Measure.* John Wiley & Sons, Hoboken, 4th edition, 2012.

[5] Patrick Bindjeme and James Allen Fill. Exact L^2-distance from the limit for QuickSort key comparisons (extended abstract). In *23rd Intern. Meeting on Probabilistic, Combinatorial, and Asymptotic Methods for the Analysis of Algorithms (AofA'12)*, Discrete Math. Theor. Comput. Sci. Proc., AQ, pages 339–348. Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2012.

[6] K. L. Chung. *A Course in Probability Theory.* Academic Press, London, 3rd edition, 2001.

[7] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition.* The MIT Press, 3rd edition, 2009.

[8] L. Devroye. Exponential bounds for the running time of a selection algorithm. *Journal of Computer and System Sciences*, 29:1–7, 1984.

[9] L. Devroye. On the probablistic worst-case time of "Find". *Algorithmica*, 31:291–303, 2001.

[10] Luc Devroye and Omar Fawzi. Simulating the Dickman distribution. *Statist. Probab. Lett.*, 80(3-4):242–247, 2010.

[11] J. A. Fill and T. Nakama. Analysis of the expected number of bit comparisons required by Quickselect. *Algorithmica*, 58:730–769, 2010.

[12] J. A. Fill and T. Nakama. Distributional convergence for the number of symbol comparisons used by QuickSelect. *Advances in Applied Probability*, 45(2):425–450, 2013.

[13] James Allen Fill. Distributional convergence for the number of symbol comparisons used by quicksort. *Ann. Appl. Probab.*, 23(3):1129–1147, 2013.

[14] James Allen Fill and Mark Lawrence Huber. Perfect simulation of Vervaat perpetuities. *Electron. J. Probab.*, 15:no. 4, 96–109, 2010.

[15] Philippe Flajolet. Singularity analysis and asymptotics of Bernoulli sums. *Theoret. Comput. Sci.*, 215(1-2):371–381, 1999.

[16] Michael Fuchs. A note on the Quicksort asymptotics. *Random Structures Algorithms*, 46(4):677–687, 2015.

[17] R. Grübel. Hoare's selection algorithm: a Markov chain approach. *Journal of Applied Probability*, 35:36–45, 1998.

[18] R. Grübel and Z. Kabluchko. A functional central limit theorem for branching random walks, almost sure weak convergence, and applications to random trees. *ArXiv e-prints*, October 2014. Available at http://arxiv.org/abs/1410.0469.

[19] R. Grübel and U. Rösler. Asymptotic distribution theory for Hoare's selection algorithm. *Advances in Applied Probability*, 28:252–269, 1996.

[20] C. A. R. Hoare. Find (algorithm 65). *Communications of the ACM*, 4:321–322, 1961.

[21] C. A. R. Hoare. Quicksort. *Comput. J.*, 5:10–15, 1962.

[22] H. Hwang and T. Tsai. Quickselect and the Dickman function. *Combinatorics, Probability and Computing*, 11:353–371, 2002.

[23] P. Kirschenhofer and H. Prodinger. Comparisons in hoare's find algorithm. *Combinatorics, Probability and Computing*, 7:111–120, 1998.

[24] D. E. Knuth. Mathematical analysis of algorithms. In *Information Processing 71 (Proceedings of IFIP Congress, Ljubljana, 1971)*, pages 19–27. North-Holland, Amsterdam, 1972.

[25] B. Kodaj and T. F. Móri. On the number of comparisons in Hoare's algorithm "FIND". *Studia Sci. Math. Hungar.*, 33(1-3):185–207, 1997.

[26] J. Lent and H. M. Mahmoud. Average-case analysis of multiple Quickselect: An algorithm for finding order statistics. *Statistics and Probability Letters*, 28:299–310, 1996.

[27] H. M. Mahmoud, R. Modarres, and R. T. Smythe. Analysis of Quickselect: An algorithm for order statistics. *RAIRO Informatique Théorique et Applications*, 29:255–276, 1995.

[28] H. M. Mahmoud and R. T. Smythe. Probabilistic analysis of multiple Quickselect. *Algorithmica*, 22:569–584, 1998.

[29] Michael Mitzenmacher and Eli Upfal. *Probability and computing.* Cambridge University Press, Cambridge, 2005. Randomized algorithms and probabilistic analysis.

[30] Ralph Neininger. Refined quicksort asymptotics. *Random Structures & Algorithms*, 46(2):346–361, 2015.

[31] H. Prodinger. Multiple Quickselect—Hoare's Find algorithm for several elements. *Information Processing Letters*, 56:123–129, 1995.

[32] M. Régnier. A limiting distribution of Quicksort. *RAIRO Informatique Théorique et Applications*, 23:335–343, 1989.

[33] Haskell P. Rosenthal. On the subspaces of L^p ($p > 2$) spanned by sequences of independent random variables. *Israel J. Math.*, 8:273–303, 1970.

[34] H. Sulzbach. On martingale tail sums for the path length in random trees. *ArXiv e-prints*, December 2014.

[35] B. Vallée, J. Clément, J. A. Fill, and P. Flajolet. The number of symbol comparisons in Quicksort and Quickselect. In S. Albers et al., editor, *36th International Colloquium on Automata, Languages and Programming (ICALP 2009), Part I, LNCS 5555*, pages 750–763. Springer–Verlag, 2009.

[36] David Williams. *Probability with martingales.* Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991.

Vita

Jason Matterer

Contact Information

Applied Mathematics and Statistics
Johns Hopkins University
3400 North Charles Street *E-mail:* jtmatterer@jhu.edu
Baltimore, MD 21218

Research Interests

Analysis of algorithms, probability, stochastic processes, random graph models, and machine learning on graphs.

Education

Johns Hopkins University, Baltimore, MD USA

- PhD., Applied Mathematics and Statistics, December 2015

- M.S., Applied Mathematics and Statistics, May 2010

University of California, Berkeley, Berkeley, CA USA

- B.A., Mathematics, May 2008

Refereed Publications

- J A Fill and J Matterer, "QuickSelect Tree Process Convergence, With an Application to Distributional Convergence for the Number of Symbol Comparisons Used by Worst-Case Find," *Combinatorics, Probability and Computing*, 23, pp 805-828, 2014. doi:10.1017/S0963548314000121.

Awards

Johns Hopkins University

- Recipient of GAANN Fellowship, 2012-2014
- Professor Joel Dean Award for Excellence in Teaching, 2014
- Departmental Excellence in Teaching, 2011

Probability and Statistics Day at UMBC

- Second Place Graduate Student Oral Presentations, 2013

Teaching Experience

Johns Hopkins University, Baltimore, MD USA

- *Teaching Assistant* **August 2009 to May 2015**

  (sample graded material and student evaluations available upon request)

- Instructed students in the areas of Matrix Analysis, Graph Theory, Financial Mathematics, Probability Theory, Stochastic Processes, Stochastic Modeling, and Introduction to Statistics
- Graded weekly or bi weekly homework assignments
- Helped prepare exams and other course material

- *Instructor* **May 2014 to Dec 2014**
  - Taught summer course in statistical analysis.
  - Developed and taught introductory statistics course for Psychology and Brain Science first-year graduate students.
- *Instructor* **July 2010 to August 2010**
  - Taught summer course in discrete mathematics.
  - Developed and presented all course lectures.
  - Wrote and graded weekly exams.